An Introduction to Machine Translation

ÉMILE DELAVENAY

THAMES AND HUDSON
LONDON

ENGLISH VERSION BY KATHARINE M. DELAVENAY AND THE AUTHOR

PRINTED IN GREAT BRITAIN BY THE CAMELOT PRESS LTD
SOUTHAMPTON

EVERYTHING that is said is said about some part of the universe of experience. . . . The universe of experience and the universe of discourse must in the final analysis, be one.

The preconceived assumption that linguistics, physics, physiology and neurology, force and energy, are all completely independent of one another is precisely what has hindered and still hinders progress, most of all progress in linguistics.

Joshua Whatmough

Acknowledgments

THE author wishes to express his thanks to the scholars and their publishers who have kindly given permission to quote from or to summarize their works. Particular thanks are due to W. N. Locke and A. D. Booth, editors of *Machine Translation of Languages* (Technology Press of the M.I.T. & John Wiley & Sons, Inc., 1955); to A. D. Booth, L. Brandwood and J. P. Cleave, authors of *Mechanical Resolution of Linguistic Problems* (Butterworths Scientific Publications, 1958); to the Academy of Sciences of the U.S.S.R. and to D. J. Panov and I. S. Muhin; and to Erwin Reifler, for permitting the use of duplicated reports and studies.

Transliteration of Cyrillic Characters

Russian names are transliterated throughout in accordance with the norm established by the International Standards Organization, which aims at transcription of characters, not pronunciation.

Contents

ILLUSTRATIONS

CHAPTER I

Translation in the Atomic Age

FROM 1954 onwards the Press has from time to time announced the invention or completion of a "translating machine". These news items have been premature, and more likely to hinder than to help research, since they tended to encourage a passive attitude towards a problem which still requires much patient investigation and the collaboration, in new fields, of specialists hitherto little accustomed to work together—linguists and electronics engineers. Now in an advanced stage of planning, and certain within a few years to become an accepted tool, the translating machine, to all intents and purposes, is already with us. We can therefore rely on the inventiveness of *homo faber* and study here, without entering the realm of science fiction, the origins, workings and potentialities of this invention.

It would no doubt be useless to swim against the stream and to call it by another name. The law of least effort will assure the success of "translating machine" by analogy with sewing machine, knitting machine, washing machine, etc., even if we were to propose a formula such as "electronic translator" or "automatic translator". Yet we are concerned not so much with a new machine as with a new analysis of linguistic phenomena, particularly of discourse, with a technology of language, made possible by the application of electronics to the signs in which thought materializes in the form of language. If we adopt here the accepted terminology and speak of the translating machine, of the automatic or electronic translator, it will be well to remind ourselves frequently that we are dealing not with a robot brain replacing the mind of man, but with a tool at the service of the human intellect and that the main effort of research, which must be primarily linguistic, will have to be focused on the process of translation, and not on the invention of a machine, i.e. an assembly of parts and electric circuits. Such machines already exist. It remains only to learn to use them for the purposes of translation. For we must guard on the one hand against the cult of cybernetics and of the electronic brain, and on the other against the complacency shown by those who, hearing of Russian or American advances in the field, imagine that all will be well if they let the engineers of these great powers finish the job and then make use of their machines.

The idea of applying the new potentialities of electronic computers to translation from one language to another has been in the air since 1946. The opportunity of subjecting the material forms of language to the analytical methods of machines capable of arithmetical and logical operations was too tempting to go long neglected. Moreover, the requirements of men in this atomic age are such that automatic translation corresponds to a real need of our time.

NEW ASPECTS OF THE TRANSLATION PROBLEM

The problem of translation, which has faced modern man ever since the Renaissance has, like many other problems, taken on new aspects in the light of the geographical shifts of power apparent at the outset of the atomic era. Vast potentials of industrial power are available to serve the political ends of great empires. The conservation and expansion of this scientific potential depends upon rapid and accurate information being made available to scientists. But, today more than ever before, scientific intelligence is impossible without translation, since the fragmentation of knowledge and the intense specialization of scientists make it extremely rare to find men with minds capable of synthesis fully cognizant with various scientific subjects and accurately and widely trained in linguistics.

Ours is not an age for learned disquisitions on “Unfaithful Beauties”, the name given in the seventeenth century to Nicolas Perrot’s translations from Greek and Latin classics in which he claimed that he had embellished and improved the originals. His enemies, who supported the cause of the inimitable superiority of ancient writers, reminded him that translations, like women, were rarely both faithful and beautiful. If we now pay any attention to this old controversy, it will be rather to try and place the problem of translation in its historical perspective from the Renaissance to our own time, to put the emphasis on the needs of science and to solve in accordance with the spirit of the age this particular aspect of the perennial dispute between Ancient and Modern.

Whether the aim of those who direct the action of today's scientists is war or peace, the welfare or the destruction of mankind, science itself needs translations of the results of contemporaneous work available in *real time* and sufficiently correct to be understood. And this is true not only of the gigantic industrial enterprises of the United States and the Soviet Union, but equally so of the less well-endowed research teams of the lesser powers. It is particularly true of the older nations of Europe, which attach particular value to their languages as living expressions of their personalities but are no longer able to develop their national scientific research on the same scale as today's industrial giants. It is significant that scientists and specialists in the documentation of science, fully aware of the new and ever-increasing need, have been the first to show interest in the problem of automatic translation. It is only fair to add that perhaps they were not, like the linguists, held in the leading strings of a historical and literary training which continues to direct the study of language towards the traces of the past rather than towards the possibilities of the future.

Our scientific age is also a nationalist age. The empires of the nineteenth century, advocates of the assimilation of the native, or at best professing the theory of political and cultural independence at a very distant date, dispensed the crumbs of western culture through the intermediary of vehicular languages introduced mainly for commercial purposes. These empires have now almost all given way to sovereign nations whose first concern is to assert their independence in every respect before recognizing their interdependence with their former masters. They want to expand teaching in the vernacular language, but at the same time lay loud claim to their share in the scientific and cultural heritage of man and proclaim their right to accede to universal culture. These young nations demand translations: school textbooks, books for teaching science and training teachers, readers for children and for newly literate adults. Already, too, they are demanding translations of the great works of world literature.

Let us make no mistake: the impassioned speeches of the former Lebanese President, Camille Chamoun, before the General Assembly of the United Nations in 1946 in favour of the translation of the great works of world culture into the languages of less privileged peoples called attention to one of the great cultural problems of our time and opened up new vistas. But though greatly broadened in scope when seen in the fresh context of the nationalism of newly independent states, this particular aspect of the translation problem has not greatly changed in essence since the time of the Renaissance. What is required is not new discussion of the old theme “Is translation possible?”, but an effort to make available more authors, both ancient and modern, in an increasing number of languages. The problem of quantity comes first: quality will come later, with increased leisure. Did we not for many years devour translations of Russian novels which were certainly more unfaithful than beautiful? And are we to think that these translations were useless?

So, without wishing to offend the classicists, the problems requiring solution today are those of quantity and of speed. The hard fact is that neither in advanced countries nor in the so-called “under-developed” countries are there enough translators to satisfy the priority needs of science and to communicate to the masses all the scientific and cultural forms of knowledge. Good translators are rare, and their work is inevitably slow. As for the others, the most useful—or least harmful—are those whose translations are faithful, if pedestrian.

Why do we not train more translators? It is true that this is an urgent task for our schools and universities. It is also a very long-term and often thankless task, and one which has so far not generally been well tackled, except in a handful of institutions where the approach has been realistic, without premature seeking after literary effects. The advent of machine translation is no reason for the schools to relax their efforts. On the contrary the machines will require the services of a great number of linguists schooled in the best methods of human translation.

Whether by coincidence or by almost prophetic foresight, the first scientists who, some twelve years ago, envisaged the possibility of using electronic computers to solve linguistic problems, were moving towards a solution of these problems. The new high-speed digital computers were still in their infancy when, in 1946, Andrew Booth suggested to Warren Weaver, Vice-President of the Rockefeller Foundation, that such machines might facilitate the work of translators. Booth himself has said that his suggestion was simply an intellectual exercise directed at finding yet another use for the new machines. [7]

SOME ASPECTS OF ELECTRONIC PROCESSING OF INFORMATION

Since that date electronic computers have become so much a reality that it is unnecessary to consider here at any length their scientific, managerial, arithmetical and logical applications. But the investigations first prompted by Booth and Weaver are part of a whole series of theoretical and practical research work which it will be useful to consider briefly before proceeding to examine in detail the methods by which written translation can be made automatic. It is also necessary to define their relationship, more or less direct as the case may be, to various scientific theories or fields of research.

In the fields of technology, of natural and social science, an effort is being made to facilitate and accelerate access to information by the use of modern methods of recording, classification and search. The first attempts were made with various types of card index, then with punched cards, whence it was relatively easy to proceed to coding methods used by computers. Three different types of system are at present in use: some based on punched cards sorted either mechanically or electronically; others using photo-electric sorting; while in magnetic recording systems, the sorting is accomplished by methods similar to those of the big computers. If the methods of analysis and indexing of scientific documents, of coding and classifying the information they contain, have not evolved rapidly enough to allow a really revolutionary and economic use of the new techniques, it is because these methods are closely related to one of the most important aspects of research on translating machines: the creation of electronic dictionaries. Progress in information retrieval by means of key words will only come with the solution of the lexical problems of machine translation, since both are closely concerned on the one hand, with the cataloguing and classification of concepts and the words expressing them and, on the other, with the technical improvement of “memories”.

For the purposes of official records, verbatim or summary, an attempt is being made to "receive sound waves in order to extract from them alphabetic information", according to Dreyfus Graf, who is working in Geneva on a "phonetograph" by means of which a typewriter will be able to record speech directly.

In combination with a translating machine, the phonetograph or some such machine enables us to envisage the possibility that one day interpreters will be replaced by an interpreting robot. But there is a long way to go before machines will be able to identify with certainty the meaning of a sentence containing homophones, to translate it, and to pronounce a corresponding sentence in another language synchronically with the reception of speech by the phonetograph. To be able to do this would suppose that a solution had been found to the problem of the matching of meaning from language to language while taking into account at least some of the shades of individual meaning intended by the speaker. Automatic interpretation involves other additional problems besides those of homophony. It can therefore be achieved only after automatic written translation has been accomplished.

WHERE TRANSLATION DIFFERS

Unlike machines designed for the transmission of spoken or written discourse—the telegraph or the cryptograph—machines designed to search for or to translate information must be able to choose from among the material at their disposal. And, at first sight, this choice seems to go much further in translation than in a simple search for information. Now no machine can exercise choice, except in accordance with precise criteria which have been determined in advance and written into its programme. The musical box or pianola of our childhood days also offered a choice, as does also that embryonic automatic translator—the tourist's Conversation Guide. Punched card sorters can regroup certain data according to pre-established programmes. They can, for instance, make a list of all customers in a certain business sharing certain characteristics, or of all a firm's employees under 25 years old, etc. It would be easy to make a machine which would "translate", that is to say which would print out or even read aloud ready-made translations of pre-selected texts chosen at random from a certain list. Such a machine may well impress sightseeers at a fair, but would serve no useful purpose. What science is concerned to achieve is a machine which, while remaining an object devoid of intelligence and of judgment, and performing a series of strictly predetermined operations, is capable of respecting certain of the original and individual characteristics of discourse and of reproducing them faithfully in another language.

The very conception of such a machine implies a thorough exploration of the relationship between thought and language. By exploring language in order to arrive at automatic translation of discourse from one language to another, we raise once more the question of the degree of freedom enjoyed by human thought, and we are forced to consider the constraints within which it operates. This is part of the eternal debate between a strictly determinist conception of fate and human nature and the belief in liberty. Here are problems of far more fundamental interest than that of asking whether or not the machine will be able to translate poetry or connotative language.

Thus the attempt to automatize translation leads us to a new conception of linguistic studies. It is no longer a question of delving into the past history of our languages, but of studying the actual behaviour of language in the expression of thought, of examining the inner dynamics of sentence creation, of the materialization of nascent thought and the different possibilities of expression offered by different languages. If we can think of language as a mould, or framework, predetermined to a very large extent, within which thought can express itself, we shall be in a better position to assign to that research with which we are here concerned its proper relation to various allied techniques.

An early observation concerned a superficial analogy between translation and cryptography. It is indeed tempting to use the word "translation" loosely whenever it is a question of transferring a message from one system of symbols to another. In this sense it is possible to say that a stenographer "translates" her signs before transcribing them into longhand. The telegraphist "translates" a telegram from latin script into morse, etc. We shall here use the word "translate" only to describe the transposition of discourse from one language to another.

Warren Weaver had not considered all the implications of his thought, and allowed himself momentarily to be seduced by a mere analogy, when he wrote, in 1947: “One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: ‘This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode’ ” [17]

For the linguist this is an over-simplification, especially if he is bi-lingual and trained in observing the divergent ways followed by his discursive faculty as each of his two languages offers different alternatives, even for the expression of precise facts and scientific data. Weaver’s idea, which made for optimism at a time when everything was still to be done, corresponds, however, only to a very elementary state of research in the field. The history of work on machine translation confirms I. K. Bel’skaja’s observation that anyone basing research on the idea that the problems of translating language by machine might be similar to those of cryptography would inevitably be disappointed. [5]

In reality, in cryptography, coding and decoding, however complex, operate always within the framework of a given linguistic structure, to which the coded message must also conform. The semantic and syntactic conventions are necessarily common to both author and reader of the message, whatever transformations the message may undergo between point of departure and point of arrival. Translation from one language to another requires something else altogether, and it is precisely this "something else" that has been the object of research during the last ten years. We shall none the less frequently have recourse to the experience of coding and decoding experts, experience which is indispensable for the realization of mechanical translation. We should not, however, mistake the part for the whole.

The same observation holds good, but requires much finer discrimination, concerning those aspects of cybernetics grouped under the head of “information theory”. It is clear that if, as G. Th. Guilbaud asserts, “one of the most active branches of cybernetics will . . . be the application of statistical methods to phenomena of which one of the dimensions is time” [14], discourse is bound to interest cybernetitians precisely because it is such a phenomenon and is susceptible of macroscopic analyses offering a striking analogy with the study of thermodynamics.

Studying on the one hand the structure and the measurable properties of information itself, information theory will analyse the measurable properties of sound waves of language, of alphabets, which are codes, and of those other codes, the systems of symbolic figures which transmit messages through the circuits of electronic computers. Still other codes which it will study are the words in a dictionary—those conventional signs representing things or ideas; inflexions which add information to the message conveyed by the uninflected forms; and syntactic rules which, in turn, give information by pin-pointing the individual meaning of words and their role in relation to other words. All the methods of measurement and analysis already applied to messages, to keyboards and signals by specialists in information theory, can be extended to the constituent elements of language and of discourse.

Information theory also studies the circuits through which information is transmitted, and the conditions governing their stability and efficiency. Translating machines, like electronic computers, are precisely assemblies of such inter-dependent, mutually controlled circuits. Whether it is a question of discourse itself, the object of research, or of the methods and techniques to be utilized in finding a solution to the problems involved, it is certain that information theory will constantly be called upon to contribute to research on mechanical translation.

It is, however, necessary to guard against the idea that the mathematical theory of the techniques of transmission can alone provide art easy solution to the problem of translation. Certainly, important parts of our work as a whole are closely bound up with the methods of mathematical analysis used by the statisticians of alphabets and signals; but we must never lose sight of the originality and individual nature of discourse that must be translated from language A into language B. The macroscopic application of the theory of probability helps us to distinguish some statistical properties of language as a system of information. But the usefulness of these statistical laws remains very limited in face of the expression of individual thought in discourse. As Panov has said: “The very nature of the problem of translation is such that individual features of the translated text cannot quite be ignored." [25, 26]

To look to information theory for the key to automatic translation would be to make a mistake similar to forgetting, when looking at Watt's governor on a steam engine, that this apparatus, which in its way "feeds back" information, would be nothing without the source of energy the application of which it regulates and controls. A translation machine will always have to deal with a *text*—the raw material, and, as it were, the source of energy; with the *methods* of transforming this text (in the elaboration of which information theory will play an important but limited role); and then with a *second text,* the result of the work of translation, an important aspect of which will be *quality;* this will be proportional to the degree of respect which it will be possible to give to the individual characteristics of the first text while changing systems of information, that is to say, passing from one language to another.

Nor can linguistics alone, even in its most modern forms, provide the required solutions. Irrespective of whether it studies language in its historical or in its structural aspects, this science has rarely attempted to make a systematic inventory of the complex relationships between a series of ideas in one language and the most faithful possible expression of the same series of ideas in another language. This kind of preoccupation has generally been left to the translator, a practitioner whose art rarely receives due appreciation. The mechanization of this art will never be possible until the inter-relationship in the expression of the same facts or ideas in different languages have been inventoried and scientifically analysed with the help of the most modern statistical methods and of the whole apparatus of information theory. Only such an analysis of linguistic data and of the behaviour of the constituent elements of language will enable us to progress and finally to achieve mechanical translation. And such analyses can hardly be made without the assistance of electronic machines capable of treating rapidly and correctly large quantities of detailed data and discovering their common characteristics. To comprehend both the potentialities and the slowness of such an undertaking, it is necessary to grasp the essential relationship between the achievement of mechanical translation and the empiric use of the very type of machine which we are trying to realize. Like other industrial revolutions, this revolution in the technique of translation carries within itself its own seed: to begin is everything.

Technological research on the transmission of information, the theoretical study of the mathematical laws of signals and of messages, studies in the psychology of language and comparative structural analyses—all these different lines of research form part of the combined operations which will result in automatic translation. To say that it is above all a question of technology, rather than of fundamental research, is not in any way to belittle the problem or the work required. It merely defines the limits of a particular field within a much wider area. One of the peculiarities of this field is that it lies somewhere between the macroscopic analysis of language studied from the point of view of quanta theory and the microscopic analysis of those individual states of consciousness out of which discourse is born.

CHAPTER II

Computers and Language

POSSIBILITIES AND LIMITATIONS OF COMPUTERS

To the arithmetical functions of mechanical calculating machines, the prototype of which is Pascal's adding machine, the new electronic computers have recently added logical functions; that is to say that they perform operations closely akin to certain actions of the human mind which, at first sight, would appear less readily mechanizable than the four arithmetical operations. Moreover, technical improvements to adding, accounting and statistical machines have resulted in a considerable increase in the possibilities of expression of these labour-saving tools, considered from the point of view of the detailed and explicit information communicable by their output organs. As these machines can now imitate an increasing number of mental operations, it is important to clarify their relationship to language, envisaged here as the faculty of expression of thought through the spoken or written word, and to discourse, the materialization of this faculty in the form of auditive or visual signs.

A factor common to all the above-mentioned machines is their ability to transform and reassemble data, whether consisting of numbers only, as in the case of adding machines, or of both numbers and alphabetical signs, as in the case of accounting or statistical machines and computers. Data (for example two numbers the sum of which is required) can be fed into the input organ of all these machines; they all perform operations, the ordered sequence of which, if several are required in succession, is called a *programme;* the result of these operations appears in the machine's output, in one form or another, whether it be on Pascal's adding machine, on the paper tape of the calculating machine, or on the typed statement of the tabulator. We shall use the term peripheral functions or organs for those which are concerned with input and output, and central functions or organs for those performing internal operations, the results of which are usually visible only in the final output.

Statistical tabulators have familiarized us with the high-speed output printer, in which the final results are typed out in immediately usable form. Direct reading of data by the machine is now rapidly becoming possible, thanks to the photoelectric cell. As for the central operations, such as the four operations of arithmetic, logical processing, or the matching of data received in the input with data stored in the memory, these have become fully automatic and extremely rapid in the electronic computer, where the substitution of inertia-free circuits for the components of the old mechanical calculating machines has made it possible to execute programmes of ever-increasing complexity and variety. But however flexible their programmes, however dizzying their speed of operation, the potentialities of all these machines are fundamentally limited by the very fact that they are machines, responding to signals, but incapable of doing more than reproducing, even while reassembling them in another form, the data fed into their input. Whereas the machine obeys signals, language is an exclusive attribute of its human operator, who interprets the results provided by the machine.

TABULATORS

Punch-card machines can, by appropriate entries typed out on a printed form, give expression to the relationship between certain figures and certain words designating objects or people. From input material they can prepare documents, invoices, statements, etc., taking into account relatively numerous and complex factors. Their work already goes a long way towards imitating certain associative functions which might have been supposed to have been a preserve of human intelligence.

Punched cards are used in these machines for three essential purposes: input of numerical information; input of alphabetical information; input of programme instructions. Punched cards also constitute a “memory”, since they are used to store in permanent form information usable at any time.

But whatever the purpose for which a punched card is used, the movements which it produces within the machine are always of the same type. The card is moved sideways, and its columns are sensed by brushes which activate electro-magnetic components by the action of electric circuits, through a switchboard which can be modified at will when it is desired to change the programme. A card can be divided into zones in accordance with a pre-established plan, the holes in each zone having a predetermined meaning distinct from other areas, so that the pattern made by the holes in the cards, controls the details of the execution of the programme. A punched card (see Figure 1) is divided into a given number

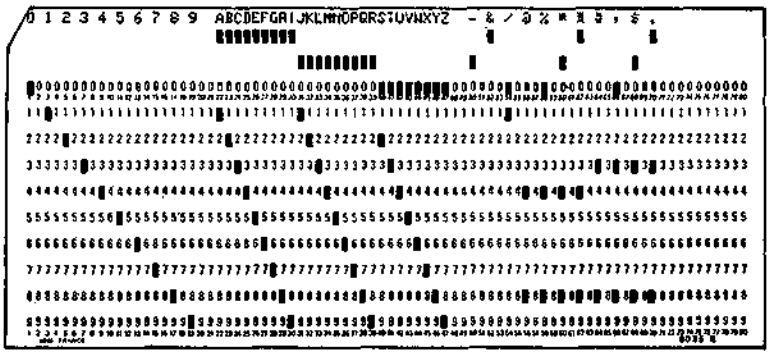


Fig. I. Sample punched card.

of—generally 80—vertical columns with ten horizontal positions numbered from 0 to 9. Other perforations can be made above these ten horizontal lines, thus allowing an even greater number of combinations. The perforations from 0 to 9 set in motion the cogs of an adding machine and the keys of the printer; the perforations at the top of the cards, in combination with positions 0 to 9, represent the letters of the alphabet and other conventional signs, and set in action either the electro-magnetic mechanisms of an electric typewriter, or any other required movement of the tabulator's mechanism.

The punched cards lend themselves to various operations such as sorting, analysing and reassembling information, and can thus be extremely useful, for instance in linguistic analysis and data retrieval. The principles involved in their use are perhaps best illustrated in the relatively simple operation of commercial invoicing. A pack of cards is fed into the machine, the first card bearing the current date and all other indications of a general nature. The second card bears the name and address of a particular client, together with the number of his account, the basic rate of discount to which he is entitled, and any other special indications applicable to this client. On other cards holes have been punched to correspond to the names of the articles ordered by this client, the unit price of each article, etc. The number of units supplied, or indications concerning units ordered but not supplied, may appear on one or several following cards. And so on for all the orders of all clients for one day. Set in motion by the cards, the machine performs all the necessary numerical operations and draws up an invoice bearing the name and address of each client, printing out in full the usual information concerning the number of articles supplied, unit price, discount, sub-totals and overall total. The invoice will indicate where necessary that such and such an article is temporarily out of stock, no longer available, etc.

The strictly automatic nature of these operations, the fact that they proceed without risk of error once the cards are correctly punched and assembled, as well as their rapidity, distinguish them from the same work performed by man. The machine has transferred alphabetic data from the cards to the printed form, in an order which is predetermined by the programme (switchboard, punching, ordering of the cards); it has performed numerical calculations, and it has transferred the results into the appropriate columns and lines of a pre-printed form. The whole effort of reflection, of intelligence, had been made during the conception of the tabulator and the establishment of its programme. The use of the machine in this particular field has thus had the effect of *drawing attention to the purely mechanical character of various operations formerly performed by a human being and accepted as mental operations*. The invention of the tabulator has pushed back the frontier between the mental and the mechanical.

Is it permissible then to speak of tabulator language? It is clear that the output component of this machine remits language in a form directly accessible to the human reader, but it does so as strictly mechanically as of a clockface which “tells” us the time.

From one end to the other of the machine's operations, signals are transmitted along wires, movements are set off by trigger actions. At the outset we find information “objectified” or "materialized" in the form of punched holes; at the end the material presentation, in the form of typewritten letters and figures, of a new combination of the data originally fed into the machine. This re-arrangement has been performed by the central organs. To speak of machine language would be almost as misleading as to speak of finger language when the hand strikes a typewriter key, on the pretext that the corresponding character, in hitting the inked ribbon, inscribes on the paper a letter which has meaning. Yet the mechanical movements have, on the one hand, performed *calculations,* and, on the other hand, imitated *association of ideas,* such as that which consists in associating the name of Mr Smith with the order for twelve coffee grinders at x cost with y discount.

Moreover, if the machine can thus imitate language and reasoning without risk of error, it is because all the numerical and alphabetical elements of its work have, like its programme, been meticulously prepared. The reasoning was done by man prior to the operation of the robot. In the punched card system and in the electric connections of the tabulator, each signal is unequivocal; choice is no longer involved. Once the holes have been punched in the cards and the cards selected and placed in the required order, we are in a world of strict conventions from which ambiguity or possibility of interpretation are excluded. The combinations of alphabetic signs which the machine transmits or reproduces have taken a unique, positive semantic value, as exact as that of the figures in the arithmetical element of the tabulator. Everything in this system is predetermined and inhuman.

HOW A TABULATOR MIGHT TRANSLATE

If we call syntax the associative faculty of the machine and vocabulary the words punched on to the cards in the form of alphabetic signs, we shall see that this mechanism is, in a limited way, able to make some sort of sentences. Evidently this same machine, with the same figures but different combinations of letters of the alphabet, can make sentences in French, in English or in German, in accordance with "instructions" provided by the alphabetic punched holes and with a programme which may, if required, alter the order of the words.

With a suitable pre-selection of cards, it might even "translate" an invoice from French into English. For this purpose a system of matched meanings would have to be established between (1) the French names of the objects designated; (2) the corresponding French alphabetic punched holes; (3) the corresponding English alphabetic punched holes; and (4) the English names of the same objects. All that would be required for instance would be a programme enabling the French alphabetic punched holes, instead of setting in motion the *typing* of the French words, to call for the corresponding English punched cards, which in turn would set in motion the typing of the English words. This would be a clumsy and scarcely useful process: nevertheless, it illustrates how a tabulator might translate—within strictly limited and entirely predetermined terms of reference. We may ask whether it is really possible to give the name of translation to this search for and typing out of signs which have a meaning in one language, mechanically reproduced by association with signs having the same meaning in another language, all within a single, predetermined syntactic mould. But does not in fact the translator do exactly this when he renders, in a list of articles on order, *one gross fountain pens* by *12 douzaines de stylos?*

Thus we are now ready to accept the idea that translations, even if only of a very elementary nature, can be mechanized with the help of a relatively simple machine to which naturally we deny any creative aptitude or faculty. The name of translation is as appropriate to these operations as is the name of arithmetic applied to the sums totalled on Pascal's adding machine. Thus the borderline between the mechanical and the mental recedes yet a little further when we observe how a machine can shift and combine signs in a manner which leads to the same results as a translation. Once we have accepted this starting point, a process of irrefutable logic will enable us to push back this limit even further as we pursue a parallel course, on the one hand, towards the creation of more complex mechanisms and, on the other, towards the analysis of discourse considered as a series of material signals meaningful to the human mind. Let us however bear in mind that we are speaking here only of thought which has already been materialized in the form of signs.

ELECTRONIC COMPUTERS

The electronic computers of today possess mechanisms sufficiently complex to permit a further and closer analysis of linguistic phenomena, not statically or in the abstract, but in relation to dynamic sequences of mechanical operations and switchings of electric circuits, the final effect of which is an output which imitates discourse. The really original character of the linguistic studies born of the research which will culminate in automatic translation is just this: discourse can now be studied in relation to the functioning of mere unconscious mechanisms, by means of the laboratory instrument provided by electric circuits.

Computers have been in existence for barely twenty years. Only since 1950 has their use become extensive enough to play a significant part in our economic and social life. While utilizing most of the old methods of mechanical computing and of tabulators, they have introduced three essentially new characteristics:

Stupendous speed of operation, resulting from the total or partial replacement of electro-mechanical cog wheels by electronic circuits so that signals travel at speeds bordering on that of light.

Increased flexibility and complexity of programmes, also made possible by electronic switching of circuits, instead of the former mechanical or electro-mechanical methods.

The extension of the central functions, logic being added to arithmetic, a development also speeded up by the use of electronic tubes, rectifier circuits and magnetic cores.

These three basic characteristics have made it possible for computers to imitate certain operations of the mind, certain mechanical aspects of which had not previously been emphasized. Simultaneously with the evolution of the central organs, the very rapid improvement of input and output media has also increased certain resemblances to human mental functions.

The first revolutionary change in computing machines was the introduction of the *memory,* that is the faculty of holding within the machine the results of a calculation before proceeding to the next one, without output of the first result and its re-input by human intervention, before the next operation is started. With the traditional adding machine it was necessary for anyone wanting to effect a series of operations to transcribe the intermediary results and then reintroduce them manually. Charles Babbage, who worked out the design of his Analytical Engine as far back as 1833, was aiming at the automatic performance of successive arithmetic operations. His machine included a memory or *store,* consisting of a group of *accumulators,* into which the results of operations made by the *mill* or arithmetic organ could be *transferred.* These partial results could also be put back into operation in the mill as and when required. The *programme,* which included *calculations* and *transfers* from memory to mill and vice versa, was controlled by two bands of punched cardboard similar to those used on a Jacquard loom. Babbage was never able to complete his machine owing to the inadequate production and tooling facilities of his time.

In 1944 the Mark I or Automatic sequence calculator of Professor Aiken followed the main outlines of Babbage's analytic machine. The inertia of its electro-magnetic relays limited both its speed of calculation and its memory capacity. The use of electronic tubes, particularly of the double triode or flip-flop, in the E.N.I.A.C. (Electronic Numerical Integrator and Calculator) of the University of Pennsylvania made it possible in 1946 to perform in 2.8 thousandths of a second a multiplication of 10 figures by 10 figures, as opposed to 6 seconds with Mark I. With E.D.S.A.C., constructed at Cambridge University, and Aiken's Mark III, the superiority of electronics was fully established; at the beginning of the fifties the great industrial and commercial enterprises began to be interested in computers. The big I.B.M. data-processers, Remington Rand's Univacs, Leo of Maison Lyons in London, Ferranti's Pegasus and Mercury, Bull's Gamma 60 in France, the B.E.S.M. in Moscow, are all endowed with high operational speeds and with arithmetical and logical components capable of extending their operations far beyond mere sequences of computation. All these machines are about to be superseded by much faster computers. Their essential organs are more or less alike.

CENTRAL ORGANS

The store or memory, as conceived by Babbage, is the faculty of holding data in reserve either permanently, e.g. a table of logarithms, or momentarily, e.g. the partial results of a sequence of operations which can be brought back into play at the desired moment in the execution of the programme. Both figures and letters can be stored in a memory, where they are represented either by holes punched on cards or on teleprinter tape, or in any other material form corresponding to the input technique employed. All that it is necessary to know here is that modern computers use different kinds of memories for different purposes which may vary in the following respects: capacity; time of access to stored information; whether data are accessible at random or according to some predetermined sequence; whether the record is permanent or not. The main types of memory now in use are magnetic tapes and discs, drums and ferrites or magnetic cores.

A memory can contain a varying number of signs or characters. On a magnetic tape made of plastic covered with magnetic oxide only ⅜ of an inch of tape is necessary to record all the information contained in the 80 columns of a punched card, so that 10,000 characters can be read in one second. The capacity of an I.B.M. magnetic tape is 5,760,000 characters. Such a tape constitutes a very high capacity memory, but with sequential, and therefore relatively slow access. The same is true on the whole of magnetic discs, which like the tapes can be arranged in batteries so that their capacity is virtually unlimited. Access-time on discs can be lowered by an increase in the number of reading heads or by high-speed motion of a single reading head.

Magnetic drums are metal cylinders covered with a magnetic product. They revolve continuously at high speed and reading and recording heads are arranged around them so as to make it possible at any given moment to record a message at a given "address" or to extract the contents of any section of the drum. These memories have a limited but considerable capacity (294.912 binary figures in the I.B.M. 704 now installed in Paris). Access to data is practically random, and is very rapid, the average access time being of the order of a few millionths of a second in the 704 and varying from 22 microseconds to under one millisecond in the Gamma 60. As with the tapes, the recordings may be preserved or erased at will.

Magnetic cores or ferrites are small rings of magnetic matter mounted in large numbers on insulating frames. An electric current passing along wire through a core creates, according to its direction, either a positive or a negative magnetic field which lasts until a new electric impulse comes to wipe out the figure so recorded. The number of figures which can be registered is limited to the number of cores, each with its pair of wires. On the other hand access time is extremely short. In the Gamma 60 it is of the order of 11 microseconds; in the I.B.M. 704 a word of 36 letters can be transferred from a ferrite memory into an operating unit or vice versa in 12 microseconds.

In all these memories data are recorded by positive or negative magnetization of surfaces or volumes. There are other possible processes, including that of photography on film or transparent disc sensed by photoelectric cells. The photoelectric disc is likely to play an important role in the linguistic and philological field as well as in information searching and abstracting of documents, particularly in the form first developed by Gilbert King at the International Telemeter Corporation. In this form of memory millions of characters and figures can be stored on a very small surface and read at extremely high speeds. Cryotrons also provide immense storage facilities on a very limited volume of matter—so that the size and speed of access of memories are rapidly ceasing to be a major preoccupation of machine constructors and users.

Other elements of a modern computer are also included in the category of *memories*—an excessively anthropomorphic term which the English language is happily able to replace by the more accurate name of *stores.* Those which we have already mentioned are in effect nothing but stores in which it takes more or less time to find what one wants, as in an index, a library or a warehouse. Others, which we might call *intermediate memories,* are designed not to store information permanently, but either to hold back its transmission in order to introduce it again at the required moment (delay lines), or to keep it during a certain stage of a sequence of operations (registers).

A memory may be used to store data, or to store programme instructions. Both take exactly the same form. The punched card of the type used in statistical machines is in fact the first memory for input data. Here the *holes* represent figures, alphabetical and other conventional signs. The first "programme memory" was the punched card of the Jacquard loom and the programme-controlling memory in Babbage's Folly was of similar type. We have seen that a programme can be controlled by the holes in punched cards, so that the same medium—the punched hole—is used both to record the data on which the machine operates, and to give the machine its instructions and dictate the order of its operations. This means that the perforation corresponding to the figure “1” or to the letter “a” may either actually represent this figure or this letter in a recorded piece of information, or it may be the material symbol of an instruction such as “multiply *x* by *y*” or “transfer the contents of the arithmetical operator into the magnetic drum”, etc. The same signal will thus mean one thing or another—be a fact to be operated on or an instruction to operate—according to its position in a sequence of signals.

THE BINARY CODE

The unification of data and programme does not end here. The universal computer, which performs arithmetical and logical operations, must be controlled by the most simple and universal impulses if excessive specialization of its organs is to be avoided. So that the input data and the programme instructions, both numerical and alphabetical, are usually communicated to the machine in a single form, that of the binary code.

We have seen figures and letters represented on cards by different combinations of the same punched holes. These same figures (from 0 to 9) and these same letters can be represented by means of a code comprising only two signs, 0 or 1 or + and −. This is the binary code, now used in most computers. A minimum of 6.41 binary signals or *bits* are required for one character of the typewriter keyboard (14); 4 bits are needed to represent the ten figures from 0 to 9; to the letters of the alphabet must be added conventional signs, punctuation, brackets, etc., so that in practice characters are fed into the memories in the form of 4 bits for figures and 8 bits for alphabetical and other conventional signs. The method used to change from decimal figures and alphabetical letters into binary digits varies according to the codes used. Figure 1 gives a concrete example of the intermediary stage. The conventional signals in this figure are transposed by the machine into binary digits, that is to say into series of 0 and 1, of positive and negative magnetizations.

There has been much talk of machine language. In reality the "language" of the central units of computers is a series of electric impulses of plus or minus value, which can effect positive or negative magnetizations of memory surfaces, corresponding to the signs 1 or 0 and susceptible of being transcribed in perforations on punched cards, or in letters and decimal figures by a high speed printer. As in the tabulator, everything which for the human card puncher was figures, punctuation signs, letters, etc., has become, in the arithmetical and logical machine, signals capable of activating mechanisms or electric impulses, of varying potentials or of magnetizing surfaces. The code has become independent of variations of meaning in the messages which it transmits. Its object is to switch circuits, to set in motion electronic or electro-magnetic movements, and each series of combinations of plus or minus, of 1 and 0 does its work in the computer with complete indifference to the fact that 0001 in binary code means 1 in the decimal system and that 00001011 may mean "a" or that this letter forms part of such and such a word.

HUMAN LANGUAGE AND MACHINE SIGNALS

While the very great speed of the computer makes it possible to utilize this universal instrument, the binary code, the principle involved is no different from the methods described a propos of the action of punched cards in the tabulator. Once again we are dealing not with *language* but with a sequence of electronic operations transformed at the *output* into signs which a human reader can turn into language by ascribing to them a meaning, that is to say by establishing a meaningful relationship between these signs and an exterior reality. Without this reference to objects, the signals are void and belong entirely to the domain of information theory and not to that of the psychology of language. It is important to remind ourselves of this fundamental difference between human language and what has been called, by extension and by analogy, machine language, since, in the first case we are dealing with a conscious mind establishing a relationship between the signal and the object it represents, and utilizing mechanical methods of acoustic or visual transmission and reception to represent it, whereas in the second case we are dealing only with transmissions which in themselves are devoid of meaning.

THE CENTRAL UNIT OF A COMPUTER

The central unit of a computer does however contain processing organs which combine input data with others stored in its memory.

The actual name of these elements of the central unit varies with the different makes of machine and with the differing conceptions of their manufacturers. In order to simplify and to avoid too many technical details we shall here restrict ourselves to those organs which answer to the needs of translating machines.

The arithmetical unit performs the four ordinary arithmetical operations. An addition or subtraction takes 180 microseconds and a multiplication between 400 and 800 microseconds. For the main purposes of translation the machine needs only to add and to subtract.

The collator makes it possible to compare two alphabetical words of varying length. This operation is performed by comparing (subtracting) the binary figures representing them, the result being zero if they are identical. Anyone accustomed to consulting a bilingual dictionary in translation work will at once appreciate the value of this operation, which consists of matching a word of the input text with a word recorded in the memory. This matching operation, when successful, can command the output of the equivalent word in another language.

The logical processer performs logical operations including the determination of appropriate programme instructions. It is essentially an organ capable of performing an instruction of the following type: if the result of a certain operation is positive or zero, execute instruction (*a*); if, on the contrary, the result of this operation is negative, execute instruction (*b*). This instruction is known as a "jump" or "conditional transfer order". If, for example, the word "works" is a noun, the machine must translate it into French by "travaux", but if it is a verb, by "il (elle) travaille". The translation programme will in such a case be controlled by a "jump" instruction—the memory containing the notations "noun" and "verb" against the word "work". The collator will match the input word against a dictionary word, then it will identify its grammatical role in the sentence and check that against the two grammatical notations. Only then will it find the

French equivalent of the word in the sentence under consideration. All the various instructions for the performance of this sequence of operations are given by the logical processer until the sub-routine for this word ends in the correct matching of the English word. In the same way any choice between possible sub-routines will be determined by the logical operations of this organ, which will take charge of selecting the appropriate sequence of programmes whenever particular circumstances require it, as often happens in computation for business management, and as will happen in translating.

A modern computer comprises various *registers* on which data are recorded during analysis or operation; some for storing changes in programme, some for storing addresses (an address being an indication of the place where a certain piece of information or instruction is to be found in a memory), some which are accumulators for storing partial results of operations, etc.

Thanks to the binary code, the computer can receive numerical and alphabetical data and instructions in a single form admirably adapted to the switching of circuits. Thanks to its central units, it can combine these data in *numerical* or *logical* operations in order to produce the results required in the solution of complex problems demanding matching of multiple data and numerous sequential calculations. It can provide the results of its work at its output in readable form. It is capable of infinitely more complex and sustained operations than those required for a simple translation from one language to another. It can perform operations of matching, identification, analytical logic and arithmetic on any kind of data, provided that these data can be conventionally expressed in figures and letters or in any other system of agreed signs.

For these machines (each word being treated as a group of alphabetical symbols) it is child's play to search a memory for the exact French equivalents of thousands of English words and to write them down one after another. But when English words make sentences, the corresponding French words rarely make French sentences, unless the machine has been given the necessary instructions for substituting French syntax for English syntax, for changing the order of certain words and for conjugating the verbs differently, for looking up signals other than those of the alphabet, etc. The most modern machines can do all this, provided they are given a programme which takes into account all alphabetical and other signals included in a written sentence. The work of the past ten years has been directed towards the achievement of just such a programme.

CHAPTER III

Variations in Approach

THE idea of automatic translation has generally been greeted by linguists and translators with a certain degree of scepticism, the natural result of their inbred knowledge of the difficulties of translation. Very few have studied the structure and content of language with the strict discipline of the natural sciences, examining them with instruments or methods equivalent to the microscope, the slow motion projector or mathematical analysis. It is scarcely surprising therefore to find that the ideas resulting from the early co-operation of linguists and electronics engineers appear on some points very far removed from what are now accepted as the main avenues of research in this field. We shall, however, be able better to understand the present state of such research if we first examine briefly the past history of these new studies, the evolution of the conceptions which underlie them, as well as of certain points of detail. Moreover, in many respects this evolution has been, and still is, dependent upon the perfecting of computers and on improvements in techniques of memory and of input. Without the hesitations and false starts of the pioneers, today's bold advances would have been impossible.

BRIEF HISTORY OF RESEARCH

From Trojanskij to 1952. It appears that the invention of the Russian Smirnov-Trojanskij, patented in Moscow in 1933, made it possible to translate Russian into several languages simultaneously over a telegraph. But Soviet linguists failed to respond when he sought their support in 1939, and the Institute of Automation and Telemechanics of the Academy of Sciences was equally unforthcoming in 1944.

In 1946 a dual approach to the problem of mechanical translation was made by the Englishman A. D. Booth, and Warren Weaver of the Rockefeller Foundation. In response to Weaver’s suggestion that wartime decoding methods might be applied to language, Booth pointed out that an electronic computer is capable of storing a sufficient quantity of data to make it possible to effect a "word-for-word" translation of the type which might be made by relying exclusively on a dictionary.

Up to this point there was no question of syntax, or of word order, nor even of translating all the words of a text. The idea was simply that, to help the scientist to understand a document in a foreign language, one might usefully put before him a translated list of keywords, relying on his intimate knowledge of his subject to enable him to find a guiding thread of meaning through the disconnected words.

At Princeton, Booth and Britten began to work out the instructions necessary to enable a computer to consult a dictionary recorded in its memory and to provide a word-for-word translation of sentences fed into it on punched tape. In 1948, Richens, another Englishman, introduced the idea of automatic grammatical analysis of word-endings. This can not only improve the translation by giving the reader information on the grammatical role of words, but also speed up the looking-up of words in the electronic dictionary, since in theory it reduces the total number of entries. Word-for-word translation can now take the following form: *amat*—*to love* 3 p.sg. *works*—(1) *travail* sb.pl. (2) *travailler* 3 p.sg. The coded grammatical indications assist the reader of the translation in understanding the meaning of groups of words.

In 1949 Weaver pointed out that by penetrating beyond the apparent divergencies from language to language, one discovers *statistical invariants,* as found in cryptography and recognized by information theory, *semantic invariants,* as observed by the sinologist Erwin Reifler between languages having no historical link, and *logical invariants,* as described by Reichenbach. These invariants, Weaver thought, may correspond to certain basic characteristics of the human brain and to the common psycho-social origins of language. Referring to the work of Booth and Richens, Weaver maintained that a purely word-for-word translation is capable of rendering great services to scientific and technical research. He also went much further and raised the question of the possible solution of semantic ambiguity by exploration of immediate context.

The logical elements of language, Weaver claimed, can be treated by the logical circuits of the computer; Shannon's information theory can throw statistical light on translation problems, particularly if studies of statistical semantics are undertaken in the light of this theory. Finally he raised the vital question of research into "the common base of human communication—the real but as yet undiscovered universal language".

As early as January 1950, Reifler circulated privately his first study on machine translation, the first serious attempt by a linguist to analyse the preparation of written texts for translation by computer. He postulated the necessity both for pre-editing texts before translation and for post-editing them when translated.

Research schemes began to multiply, with apparent lack of co-ordination. Oswald and Fletcher studied the mechanical resolution of German syntax patterns. At the Massachusetts Institute of Technology Bar Hillel, the Israeli logician, was able to devote his whole time to research on language with a view to mechanical translation. Early in 1952 the Rockefeller Foundation made it financially possible for M.I.T. to call the first conference of linguists and electronics engineers devoted to mechanical translation. The eighteen participants agreed on the next two stages of research. The first step was to undertake, in scientific texts, studies of word frequency and language-to-language equivalence, while analysing the methods for using electronic memories and examining other technical aspects of the automatic dictionary. Later would come syntactic analysis for the elaboration of machine translation programmes. The study of the circuits necessary for the resolution of grammatical and syntactic problems could be left until a later stage. Work on a multi-lingual machine should await the first results of one-way translation from language A to language B. But the possibility of using an intermediary language —*machinese*—capable of serving as a turntable between all languages, was not excluded.

1952-1955. American research developed rapidly as a result of this first exchange of views. Studies were made of the storage capacity of memories, the usefulness of restricted vocabularies, the mechanical identification of meaning and of word-endings, etc. In 1954, Dostert and Garvin of Georgetown University and Sheridan of I.B.M., successfully carried out on a 701 I.B.M.

computer the first experiment in automatic translation from Russian into English with a vocabulary of 250 words and six syntactic rules. Also in 1954, M.I.T. published under the direction of William Locke and Victor Yngve, the first number of a periodical entitled *M. T. (Mechanical Translation).* In the following year the first published book on the subject appeared—*Machine Translation of Languages,* edited by William Locke and A. D. Booth.

The year 1955 also brought the first news of Russian activity in the field. The big B.E.S.M. computer of the Institute of Precise Mechanics and Computer Technology of the Academy of Sciences of the U.S.S.R. was used for experiments shortly to lead the Academy to the conclusion that automatic translation was possible.

The expansion of research. In October 1956, M.I.T. called the first international conference on machine translation, at which foregathered some thirty specialists from Great Britain, Canada and the United States. Dr D. J. Panov, of the Institute of Precise Mechanics of the Academy of Sciences of the U.S.S.R., sent an important written contribution on Russian research. The three main centres of activity were Great Britain, the United States and the Soviet Union; smaller groups were at work in Italy and Scandinavia. The problem was no longer to prove by preliminary research that mechanical translation was a possibility, but to organize this research in such a way that effort was concentrated for maximum efficiency and that synthesis of scattered studies should become possible.

By 1959 a dozen or more groups were working actively in the United States. At Harvard, Oettinger continued and expanded his work on the Russian-English automatic dictionary. At M.I.T., Locke, Yngve, Chomsky and others were studying syntactic structures, German syntax, basic methodology and devising methods enabling the linguist to programme for a computer. At Georgetown, Garvin and Zarechnak concentrated their efforts on Russian syntax, while A. F. R. Brown worked on translation from French. At the University of Michigan, Koutsoudas and Korfhage were at work on Russian, with particular emphasis on the polysemantic problem. In Seattle, at Washington State University, Reifler, Micklesen, Wall and Hill were working mainly on Russian, in collaboration with the International Telemeter Corporation of

Los Angeles, where Gilbert King had designed a photoscopic memory, with high capacity and very rapid access, now being further developed at the Rome (N.Y.) Air Force I.B.M. research station. Everywhere translation from Russian into English had high priority. Russian/English research groups were at work at the University of California, at the California Institute of Technology, at the Ramo-Wooldridge Corporation of Los Angeles, and at the Rand Corporation of Santa Monica, where Hays and Harper had undertaken a most interesting series of studies in methodology and were seeking to synthesize work already done. The National Science Foundation of Washington was financing much of this research and endeavouring to co-ordinate it.

In Great Britain, Booth, Brandwood and Cleave, at Birkbeck College, with the financial help of the Nuffield Foundation, were also studying the methodology of research and doing practical work on Braille, French and German, much of which is described in their book *Mechanical Resolution of Linguistic Problems* [7]—a rich source of facts and ideas. The Cambridge Language Research Unit, under the direction of Margaret Masterman, was engaged in studies of lexicography and universal syntax and attempting to apply to lexis and syntax the idea of a “mechanical thesaurus”, while Richens explored the possibility of an algebraic type of universal language.

In Moscow, centred round the Academy of Sciences of the U.S.S.R. a more vigorous concentration of talent and effort appears to have been achieved, although later evidence suggests the development of conflicting schools of thought between the various groups at work. In September 1956, the review *Voprosy Jazykoznanija (Linguistic Problems)* began to devote regular space to the problems of automatic translation—with particular emphasis on the work of the Steklov Mathematical Institute and the University of Leningrad. The more empirical school, that of Panov and Bel’skaja, both working at the Institute of Precise Mechanics and Computer Technology of the Academy of Sciences, defined a general methodology and guiding principles for the collaboration of linguists and electronics engineers. Korolev was working with this group on the problems of code compression for dictionary making, and Razumovskij on the automatic programming of translation machines. Mel’čuk and O. S. Kulagina, collaborators of Ljapunov at the Steklov Institute, published a most interesting study on translation from French into Russian. K. T. Mološnaja, at the same Institute, demonstrated how the work of Jespersen and Fries on structural linguistics can be used for the resolution of syntactic differences between languages. Fifty-six papers were presented to an important scientific and technical conference held in Moscow in May 1957, among them contributions on the automatic translation of German, English, Hungarian and Chinese, as well as on experiments in translation from French. The Conference noted the need for directing linguistic studies so that linguistics might be treated as a natural science, making extensive use of mathematical methods of analysis.

A year later, another conference on automatic translation resulted in the publication of abstracts of seventy-one contributions on the subject [28]. Again news has been received of a Leningrad conference held in April 1959 where 59 papers were read [28a]. It is clear that applied and mathematical linguistics are being studied in the U.S.S.R. with new vigour and enthusiasm and with noteworthy clashes of theoretical views.

THE EVOLUTION OF IDEAS

While the basic ideas leading to automatic translation have not changed over the course of the years, they have greatly increased in boldness of conception. The initial notion of a mere *automatic dictionary* has given way to that of completely automatic, grammatically correct translation. This evolution is due mainly to the rapid improvement in computer techniques, and to the systematic analysis of language, which for the first time has been conducted with completely objective methods and based on the potentialities and also on the limitations of computer operation.

The need to consult an electronic dictionary being an essential feature of all mechanical translation, it was natural that early research should be concentrated on the principles and preparation of such dictionaries, and on the automatic retrieval of a word in language B which is equivalent to a given word in language A.

Automatic dictionary and signalization of meaning. Reduced to its simplest expression, the automatic dictionary instantaneously supplies, for a word in language A, one or more equivalents in language B, by a simple operation of retrieval. The word in language A is input and compared with a list of words stored in the memory, i.e. the dictionary. When the signs representing one of the dictionary words coincide with those of the input word (i.e. in existing machines, when the result of the subtraction of the two figures representing these two words is equal to zero) the computer is instructed to print out the letters of the word in language B which translates the word in language A.

If the word has only one meaning, or if all possible meanings coincide with all the possible meanings of the word in language B, the "match" is perfect, and the semantic unit of language A representing the input word will immediately be represented at output by the equivalent semantic unit in language B. But whenever a word has several meanings it is necessary either to output several alternative translations, from among which the reader must try to make his choice, or else some method must be devised of discriminating from among the various meanings, in order to make it possible for the computer to print at output the particular meaning chosen. This necessitates a more complex programme for the computer, which must be supplied with instructions based on criteria for selection from among the several possible translations.

The machine can only embark on a sub-routine of this type if it receives a *signal* which will start the execution of a new instruction. Must we have recourse here to human intervention to provide this signal (in which case the operation is no longer automatic) or can some objective element contained either in the signs of the written language or in the structure of the sentence call forth the necessary order to start the sub-routine? At this point it became essential to investigate the whole problem of signalling systems in written language, which, while seemingly adequate for human communication, at first appeared very incomplete by machine standards.

The problem of the signalization of meaning is by no means simple: not only is the word *dog* a multi-meaning semantic unit—or to put it another way, several semantic units; it also contains indications of *grammatical value.* If the machine can identify *dogs* as being either the plural noun or the third person singular of the present indicative of the verb, then the absence of the *s* in *dog* is also a signal, indicating either the singular noun, or the other persons of the present indicative, or the infinitive, etc. Other objective criteria, such as the presence or absence of *to,* of a subject or article, enable us to determine whether we are dealing with a noun or a verb, and if the word is identified as a noun, then the absence of the s is a positive indication of the singular. Except in cases of exceptional ambiguity, the human mind grasps the meaning of a sentence by instantaneous interpretation of signals of this kind, without consciously recognizing them as signals at all. The machine, on the contrary, needs to identify them without possibility of error.

Thus the question arose, how to complete the automatic bilingual dictionary by incorporating into it, in such a way that they could be recognized by the machine, all those signals which make it possible for us subconsciously to identify with exactitude the grammatical value of the word *dog* in a given English sentence?

The separation of affixes. Starting from the idea that syntax was of minor importance in the understanding of a language, Booth and Richens regarded the mechanical dictionary simply as a catalogue of all the invariant semantic units—stems and affixes—in language A accompanied by their equivalents in language B. By a semantic unit must here be understood any linguistic element—whether it be part of a word, a whole word or a group of words—having a distinct meaning. Booth and Richens decided to separate stems and endings in their dictionary entries.

The Latin *am*—thus represents the idea of loving, the conjugation being rendered by the addition of the verbal endings to this stem. *Rego* will be represented by three stems, *reg-, rex-* and *rect-*; the general rule being that when the derivatives of a word are not formed by simply adding the affixes, the stem of each derivative must be entered separately in the dictionary [17]. In German, for example, *Bruder* and *Brüder* will be entered separately.

If the equivalent of an input word does not figure in the dictionary, the machine searches for the longest segment of this word which corresponds exactly to an entry in the dictionary. Entries are compared from left to right, and this comparison is repeated after the identification of a first segment until all the elements of a given word have been identified. The Spanish word *comprarlo* would, for instance be decomposed as follows:

compr-	buy
-ar-	(infinitive)
-lo-	the/it

This method does not solve all the problems of semantic units composed of of several words, such as the French *ne . . . pas, ne . . . que* or German disjunctive verbs. But the use of the magnetic drum memory first made possible one solution put forward by Booth and Richens which has now become standard practice: the machine is instructed to translate the first part of a two-term semantic unit only when it comes to the second part.

The results of the trial translations of Booth and Richens were of a character likely to discourage the interest of linguists and to suggest that mechanical translation experiments would lead only to rudimentary and disappointing results. Nevertheless, Booth and Richens, by separating stems and affixes, had laid the foundations of a sound and sure method which will certainly be necessary as long as the size and speed of memories play a preponderant role in the economics of machine translation.

By so doing they established by implication a rule which has proved of great importance in the study of language for automatic translation: the practical needs of programme-making, rather than scientific and historical norms, were their guide in separating *affixes* from *stems*. In other words, the determination of the dictionary stem, or base, was made without regard for historical linguistics: it was a question not of pure, but of applied science.

Birth and death of the "pre-editor". Reifler, like all the pioneers, was at first convinced that a human operator will have to participate actively in the work of the translating machine; human intervention consisting in improving and supplementing the signals contained in the alphabet and in written language. He first defined with precision (though after a more profound analysis of linguistic data, he later withdrew from this position) the idea of pre-editing texts to facilitate the work of the machine and of post-editing them after translation to facilitate reading.

We have just seen that the conventional signals of the alphabet and punctuation do not explicitly represent all the linguistic values of which the speaker or reader is nevertheless conscious. For instance, the word *enfant* is recognized as singular only by the absence of any signal: its gender is not represented by any sign, except, in certain cases, by an agreement of article, adjective or participle: "*l'enfant que j'ai rencontré*" as opposed to "*rencontré*". Moreover the signalling system of language A does not correspond to that of language B, nor do the omissions of the two systems coincide. In order to translate, it is necessary to compensate for the omission of signals, wherever this is required by the given pair of languages A/B. Reifler suggested the arbitrary creation, for the machine, of distinctive graphic signals supplementing those of ordinary written language. These signals were to be inserted in the text for translation by a pre-editor.

Thus the role of the pre-editor would be to provide the machine with texts explicit from the graphic-semantic point of view. Reifler even considered the idea of a supplementary spelling system which would give to both machine and reader all the signals necessary for the complete understanding of a translated text.

The problem of complementary signalization arose at two levels: grammatical—the signalization of the grammatical value of polyvalent words—and non-grammatical—the signalization of the semantic meaning of polysemantic words. It was also influenced by the restricted possibilities offered by the electronic computers existing in 1952. But the first M.I.T. conference was soon to suggest that new machines would shortly make it possible to extract from conventional writing, without complementary signalization, all essential grammatical information. It remained to be seen how they could determine the grammatical role of the constituent parts of extemporized compounds in a language such as German. If this problem could be solved, then it seemed possible that translation could become completely automatic.

German Compounds. Reifler therefore set to work on German substantive compounds, and in August 1952 he announced the demise of the German pre-editor, who had now become superfluous.

The difficulties encountered were of several kinds. In the first place the meaning of a compound word does not always depend on its constituent parts. In such cases the solution must be to list the compound separately in the dictionary together with its translation. The second difficulty was christened "the *x* factor".

In compound nouns, a letter or a group of letters may belong to one or to the other constituent parts of the word. How can the machine identify the constituent elements of *Wachtraum* in order to decompose it either into *wacht-raum,* guardroom, or into *wach-traum,* day-dreaming, the *t* being the *x* factor?

After classifying German substantives according to whether they can form right- or left-bound compounds or both, by taking as characteristic signals the space which separates the words, or the absence of such space, as well as the initial capital letter of nouns, Reifler observed that the formation of German substantive compounds is governed by rules of such a nature that a maximum of four checks with the dictionary is sufficient to identify with certainty the grammatical role of the constituents of a compound, in spite of the *x* factor.

Reifler's work on compounds was completed by the elaboration of a form-class filtering system of German words, based on some of the ideas of the structural linguists and on the use of separate memories for each form-class: four big magnetic drums contained the four main word classes and ten less important classes were placed on smaller drums. This was the first detailed classification of the grammatical categories of a language made from the point of view of the operation of a translating machine. New computer techniques will modify the material basis of Reifler's system, but its linguistic basis is permanent and well adapted to computer work.

Better than word-for-word translation. Reifler's work brought mechanical translation well beyond the point of simple word-for-word dictionary translation; as soon as an attempt was made to analyse *the relationships between words* in view of mechanization, the translation of word groups became possible. The upholders of *word-for-word translation* gave way to those advocating *phrase-by-phrase translation,* that is, by units of meaning, and not by dictionary words.

The analysis of words from the viewpoint of the operations of a mindless machine had given rise to a clarification of the role of words in the communication of the idea to be expressed: certain words or forms have a clear and precise meaning quite independent of any context—for instance *wegen* in *wegen dieser Schüler* (because of these pupils). Other forms have multiple grammatical and non-grammatical meanings, and the reader must choose the meaning appropriate to a given context on the basis of the information provided by the form of neighbouring words. The meaning of one form may be pin-pointed by another form, or there may be mutual pin-pointing of two forms. In the example given above, *dieser* may be a nominative masculine singular, a genitive or dative feminine singular or a genitive plural; *Schüler* may be a nominative, dative or accusative singular, or a nominative, genitive or accusative plural. One has four possible functions, the other six. Taken together they pin-point each other's meaning since only two common functions remain: nominative singular or genitive plural. *Wegen,* which governs the genitive, excludes any possibility other than the genitive plural. This leads to the formulation of a theory of the creation of sub-routines for the exploration of immediate context, enabling the machine to find the right translation for a word of multiple meaning or function.

In theory machine translation had by then passed the crucial point where mere word-for-word equivalence gave way to a partial rendering of the relationships between words, and to an exploration of the modification of one word by another and the mutual pin-pointing of meaning.

The Georgetown-I.B.M. experiment. Dostert and Garvin at Georgetown University were working in the same direction when they set up their 1954 experiment in collaboration with I.B.M. The interest of this experiment, prepared "manually" first on typewritten and subsequently on punched cards, is today mainly historic.

The scope of the project was strictly limited: a vocabulary of 250 words selected in various fields—general, technical, scientific, military and political. In the Russian/English dictionary a distinction was made between the indivisible Russian words and those divisible into stem and ending, the stems and affixes being stored in separate memories. The number of possible English equivalents for each Russian element was limited to two, so that when a choice was necessary for translation, that choice lay between two possibilities only.

Certain diacritical signs were added to the alphabetic coding of the letters representing the words. "Programme Initiating Diacritics" (P.I.D.)—bringing into play one of the six rules of syntax; "Choice Determining Diacritics" instructing the machine to effect a reference forward or backwards (C.D.D. 1 and 2) from the word under examination in order to search for the necessary signal to determine the choice between two translations; a third group "Address Diacritics" (A.D.D. 1 and 2) gave the dictionary address of the English equivalents associated with P.I.D.'s and C.D.D.'s

The six rules of syntax were briefly as follows:

Operation 0: Immediate translation of the given input word.
Operation 1: Reverse word order.
Operation 2: Choice dependent on a following word.
Operation 3: Choice dependent on a preceding word.
Operation 4: Omission of a word redundant in English.
Operation 5: Insertion of a word necessary in English.

The success of this experiment in automatic translation of complete sentences without pre-editing or revision, compelled attention and achieved widespread recognition of the progress already accomplished in research on mechanical translation. There was now no doubt that the automatic dictionary and the stammering translations of the beginnings had been left far behind. No doubt either but that the aim—the completely automatic translation of any scientific or technical text—was still far from being attained. Only at the cost of an enormous effort of coding and programming had the I.B.M. 701 data-processing machine been able to translate these 200 sentences at the rate of six or seven seconds a sentence.

The experiment also drew attention to the importance of linguistic problems in automatic translation. Up to this time the general belief had been that only a few science-fiction enthusiasts could be interested in a game of no possible value to linguists. Proof was now forthcoming that, given a series of sentences in one language, an electronic computer could print out a series of sentences of equivalent meaning in another language. If the machine was capable of doing this, then it was necessary to face resolutely up to the basic problems of studying language with a view to making use of the new potentialities of the machine.

1955—The turning point. Such study was almost at once taken up by the mathematical specialists, electronicians and linguists of the Academy of Sciences of the U.S.S.R., while in England, Booth was able to enlist the support of the Nuffield Foundation for his research. It may be said that 1955 put the problem well on the road to actual solution, with the accent firmly placed on the study of language, and considerable progress made in the direction of machine exploration of the constituent elements of the sentence. Word-for-word translation is still considered useful, but it is definitely out-of-date.

In the purely technical field, memories and reading units were evolving rapidly. In an essay in *Machine Translation of Languages* [17] Booth emphasized that reading speeds of 100,000 letters per second made it possible to find a word in the permanent dictionary in 1/50th of a second, and, with the introduction of ferrites, offering new possibilities in the form of temporary memories, only 1/100,000th to 1/1,000th of a second would be required.

Booth described the main characteristics of a translating machine, as follows:

(a) An input, either in the form of a "reader" of original typescript, or in the form of a magnetic tape reader.

(b) A rudimentary "computer". This need only be capable of subtracting, shifting letter patterns, recording results in storage, and discriminating on the size of numbers.

(c) A small-capacity, reasonably high-speed computing storage. This might be realized on a magnetic drum, but would more probably take the form of a ferrite matrix. Experience suggested that a capacity of 64 words, each of up to 12 letters, would be adequate.

(d) The main dictionary and grammar storage. A large magnetic or photographic drum would probably prove suitable if the micro-glossary technique were used. The capacity of this organ might be 10,000 words of 12 letters.

(e) From 4 to 28 tape feeds to input the various microglossaries in the machine repertoire. Photographic film seemed the most promising material for this.

(f) A single output magnetic tape. Only one is needed since this medium can match the speed of the computer itself.

Booth estimated the cost of this machine, which would occupy a floor space of 10 to 20 sq. ft., at roughly 100,000 dollars—probably an under-estimate.

THE CONCRETE ANALYSIS OF LINGUISTIC DATA

The methods of Reifler, Booth and of the Russian school of Panov and the Institute of Precise Mechanics are roughly convergent: all of them base their work on the idea that the special requirements of the treatment of linguistic data by computers provide an instrument for linguistic analysis which enriches our knowledge of language, an instrument capable of exploring the differences existing between the systems of expression of two languages. Theories are discarded in favour of an examination of linguistic material in its complex relationship with the expression of ideas, of an extension and development of the methods of analysis used by information theory, involving penetration in depth, below the level of the alphabet, down to the semantic and syntactic elements of language, studied from the angle of the reciprocal behaviour of two languages in process of translation.

If language A expresses an idea by certain means, and language B by some other means, what are the rules determining the mutual behaviour of these two systems? How can one formulate these rules in such a way that a machine can apply them automatically to convey in language B an idea expressed in language A? In order to do this it will obviously be necessary first to draw up an inventory of all the similarities and all the dissimilarities between the two systems of expression, and then to submit this inventory to a further analysis in order to enable the machine to use the results of the inventory, for instance, in the form of binary numbers. It is clear that this inventory and this methodology constitute the point towards which the preliminary studies of the years up to 1955 converge. This does not mean that very considerable work of mathematical analysis will not be necessary for translation programming, but this work can only usefully be done after the varying methods used by two languages to give expression to the same ideas have been fully inventoried and processed.

The Panov school in the U.S.S.R. sees in the problems of mechanical translation one aspect of a group of questions to which insufficient attention has been given by specialists of information theory, who must learn to take into account the individual qualities of the information communicated instead of being interested only in its statistical characteristics [26]. Translation being an art in which perception of individual factors plays an essential role, a whole vast new area lies open for investigation, beyond the statistical analysis of the conventional elements of the alphabet and other signals; it is necessary to explore the actual means of expression in all their aspects, alphabetic, morphological, syntactic and semantic. Panov, recalling the highly promising work of Jespersen and Fries [13], nevertheless emphasizes the limits of structural linguistics as defined by these two authors; he observes that the logical analysis of language cannot by itself provide a solution to the problem of translation. As for the possibility of reducing the structure of language to mathematical formulae—another very tempting prospect—Panov writes: "Should it be achieved, the problem of automatic translation would join as equal those profound problems united under the name of theory of information. Unfortunately, so it seems to me, we must refrain from this tempting road. *The very nature of the problem of translation is such that individual features of the translated text cannot quite be ignored.*"

These individual features are found in the lexical content of the text. This fact, writes Panov, should govern our choice of methods: "I believe here we are faced with a problem which, though statistical in character, requires methods of analysis, similar to the experimental methods used in the study of natural phenomena." The Soviet mathematician concludes that "both lexical meaning and grammatical characteristics of the word can and should be considered in translating languages" and that "it would be highly impractical to decline the information which can be thus obtained." In the analysis of the text to be translated, he refuses to separate completely lexis, morphology and syntax, since all three elements contribute to determine the meaning of the text.

The basic principles of early Soviet research. Soviet research started in the Institute of Precise Mechanics and Computer Technology, on Academician Lebedev's B.E.S.M. computer. Almost simultaneously Ljapunov, Mel'čuk and Kulagina worked at the Steklov Institute of Mathematics on a smaller computer, the STRELA, and followed slightly different methods, more directly inspired by the structuralists. The efforts of the Panov group were directed less towards a theoretical comprehension of the general problem of machine translation than towards a detailed investigation of lexical material. The Steklov Institute group on the other hand were concerned with profound theoretical research in the area of mathematics and linguistics, and saw mechanical translation as part of the larger problem of the automation of thought processes. They endeavoured in particular to determine the correspondence between the grammatical structures of two given languages. The work of Ljapunov, aiming at gradual automation of the whole process of machine translation, inspired the research of several other groups: in the Institute of Linguistics of the U.S.S.R., universal rules for the analysis and synthesis of a text are being worked out by Mel'čuk, and at the Experimental Laboratory of Machine Translation of Leningrad State University, N. D. Andreev is also working on some abstract logical system capable of serving as an intermediary language, constructed by averaging the phenomena of various languages.

Yet despite some marked and sometimes even vehement expressions of divergence as to theoretical approach, the Soviet scientists and linguists are in the main following certain practical principles which were developed in 1956 by the Panov group—the more empirically minded and also until now the only one which has actually achieved limited but genuine translation by computer techniques. Those principles are probably fundamental to all practical achievement and may hold the key to future progress.

From the very beginning of their work on machine translation the workers of the Institute of Precise Mechanics decided to organize their research on bases quite different from those of the Georgetown-I.B.M. experiment. "In our opinion," writes Panov, "excessive contact between the translation programme and the ascription of the control codes directly to the words in the dictionary cannot but limit the possibilities of translation, making the solution of the problem extremely complicated. Therefore we made it our point to work out basic principles of machine translation before starting. Our five basic principles are the following:

1. Maximum separation of the dictionary from the translation programme. This enables us easily to enlarge the dictionary without changing the programme.
2. Division of the translation programme into two independent parts: analysis of the foreign language sentence, and synthesis of the corresponding Russian sentence. This enables us to

utilize the same Russian synthesis programme in translation from any languages.

3. Storing all the words in the dictionary in their basic form. This enables us to make use of the standard Russian grammar in the synthesis of Russian words.
4. Storing in the dictionary a set of invariant grammatical characteristics of a word.
5. Determination of multiple meaning of the words from the context whereas their variant grammatical characteristics are defined by analysing the grammatical structure of the sentence.

These principles have proved quite reliable in the practical test they were put to, and hence they must be considered as basic in the solution of the problem." [26]

Dr Panov has not deviated from his opinion that these five principles will lead to a complete solution of the problem. It does indeed seem that they may be able to serve as the basis for the greater part of research, whatever pair of languages are under consideration and whatever the differences between them—whether they be languages rich in inflexions such as Russian or poor in inflexion and rich in structural meaning such as English. Details of the programme will vary, but the application of these principles will remain the basis of research of a very flexible nature, and will avoid much exploration of blind alleys.

To Booth and Richens, to Reifler and the M.I.T. team, to Panov and his colleagues, must go the main credit for clearing the ground and laying the solid foundations upon which it is now possible to build.

CHAPTER IV

From Source Language to Target Language

INVENTORY OF MEANS OF EXPRESSION

To translate from a given language A into a given language B is to attempt to reconstitute with the system of expression of language B, the meaning of a sentence or string of sentences, expressed in language A by means of the system of expression peculiar to that language. The meaning of a sentence is the representation in the speaker's mind, materialized by means of phonetic and visual symbols grouped into words. Each word possesses, or may possess, several values, semantic or grammatical. Each word may be syntactically associated with other words in a number of ways. Perception of meaning is dependent on the determination of these different values and associations. Translation becomes possible only after an analysis of all the linguistic elements of language A, or source language, constituting meaning, embodied in the words and in the relations between words, i.e. semantic values, grammatical values (whether expressed by inflexions or otherwise) and syntactic values.

This analysis is followed by a synthesis of the linguistic elements of language B, or target language, selected because they make it possible to render approximately the same meaning as the original sentence in language A, and combined according to the rules peculiar to language B.

The experienced human translator performs this operation with a degree of difficulty which varies with the clarity of the text for translation and also with the degree of similarity between the structures and semantic content of languages A and B. Before the translator can be replaced by a machine, it will be necessary to prepare all the operations of analysis, of synthesis and of enumeration required for the elaboration of programmes and sub-routines enabling the electronic machine to transpose the content of a text in language A into a text in language B. The principal contribution of the linguist to the preparation of programmes and sub-routines will consist in making an inventory of all the means of expression of languages A and B, and of the relationships which can be established between their respective systems of expression.

The linguist will endeavour to draw up this inventory in such a way that it is easily reducible to numerical data and coded instructions for use by the machine. Clearly, the more closely the means of expression of the two languages resemble one another, the simpler the programme. The greater the difference between the structures, the morphology and the semantics of the two languages, the more numerous the ramifications of the programme or sub-routines required, that is to say the more circuitous the ways of finding exact equivalents between a word in language A and a word in language B. It is therefore true to say that programme economy will depend directly on the degree of structural relationship between the two languages under consideration.

LANGUAGES, INTERLANGUAGE AND METALANGUAGE

Considerations of economy, as well as certain preconceived ideas of a philosophical and logical nature, have given rise to a somewhat premature discussion concerning an intermediary language, or interlanguage, the use of which, it is alleged, might facilitate automatic translation. The desire to find such a philosopher's stone of machine translation appears to have been enhanced by the early absence of any firmly established basis for more clearly rewarding empirical research.

The issues of this debate have not been clarified by the inevitable talk of "machine language". And confusion has only been increased by the idea that the binary code controlling the actions of computers is itself a *language.* In a field in which language and languages are both the subject and the instrument of research, confusion due to purely verbal analogies and to the metaphorical use of the word language is very frequent. Fortunately Andreev, a Russian, and Mounin, a Frenchman, have, each in his own fashion, contributed to a better understanding of this complex subject in which the utmost exactitude in terminology is essential.

Georges Mounin rightly distinguishes between pseudo-languages—of which Esperanto is the classic example—intended to be *speakable,* and interlanguages, designed for use as auxiliary languages, such as the *interlingua* of Peano or that of Gode and Blair. He points out that the present work on machine translation must lead to the study of the problem of languages "which can be used as common intermediaries, as central links in the chain of translation from any language into any other". [21]

Here we must make a distinction between two ideas: between that of an *interlanguage* set up *a posteriori* as the result of research undertaken on a number of pairs of languages analysed in view of automatic translation from one to another—which would in fact express the highest common denominator of the means of expression of all these languages—and that of an *interlanguage* conceived *a priori* as a universal translation programme applicable to all languages. Whereas the first would be an outcome, the second would be a starting point. This means that before designing and working on translating machines it would be necessary first to draw up a universal programme. Is such a programme really desirable? Is this *a priori* possible?

Booth, Brandwood and Cleave [7] have demonstrated the weakness of the economic arguments invoked in favour of the universal programme thesis. For N languages, we are told, we should need N-1 programmes in order to translate, without an intermediary language, from each one of these languages into another. And to translate N languages into N-1 languages, we should need N (N-1) programmes, i.e. almost N^2. On the other hand, it is maintained, the use of an artificial intermediate language M would require only N programmes to pass from N languages into this language M and as many again to go from M into N languages, i.e. 2N programmes in all instead of N^2.

Booth easily refutes this by a mathematical argument. If the turntable language selected is not an invented language M, but a real language L1, we should need not 2N but 2N—2 (i.e. two less) programmes to translate, via language L1, all languages into all other languages. Thus a natural language would be a more economical turntable than an artificial one.

Booth also puts forward arguments founded on common sense and observation. We are told that the artificial language would facilitate the passage from language A into language B if the two languages are very dissimilar in structure. If we examine this contention closely, we shall see that language M, the entirely artificially-created language, does not in fact simplify anything. The elaboration of a translation programme A→B represents a sum of work less great than that of a double programme A→M and M→B. Moreover, if, as is proved by observation, programmes are not in fact reversible, the establishment of programmes A→M→B and B→M→A would be more costly than that of programmes A→B and B→A.

Thus we are brought back to the comparative and empirical study of languages A and B as the only practical method of setting up translation programmes A→B and B→A. And it is certainly true that if ever an intermediary language (or a universal programme of translation) becomes possible, this intermediary language is more likely to be of real use if its structure and its characteristics are experimentally based on the comparative study of multiple bilateral programmes of the type A→B and B→A.

Both the British and the Russians are, however, moving towards the use of their own languages as natural turntables, justifying their choice by various practical considerations. In spite of certain objections due to the phonetic and structural ambiguities of English, Booth sees in his own language a possible pivot, even if this should involve slight modifications of current English in the interests of its universal use. Panov, for his part, envisages the use of Russian as the basic language. Andreev [1] employs very strong theoretical arguments to justify the use of a language closely related to Russian. Leaving aside all questions of national ambition, can this coincidence be purely fortuitous? When the work of analysis of one or more foreign languages has been successfully completed, is not each national team naturally inclined to seek the general solution *most convenient to itself?* This solution naturally consists in using as turntable that pre-linguistic state prior to output when linguistic data are still expressed in code and are on the point of being transformed into English for some and into Russian for others. So that we are forced to admit, leaving aside for the moment Andreev's theoretical arguments, that other natural claims to the title of turntable might equally well be put forward. In point of fact any language spoken by a nation dynamic enough to carry out a number of bilateral translation programmes into its own language can also claim to be a natural turntable language. As we shall see below, however, the optimum choice may depend on certain factors inherent in the structure of this language.

While an intermediary translation machine language remains, for the moment, if not a somewhat academic dream, at best a very long-term project, this is not true of a *metalanguage* for the use of specialists in automatic translation. The automatic transposition of language A into language B can only be performed in accordance with a strict system of instructions given to a machine, which responds to numerical codes representing the letters of words, grammatical forms and syntactic relationships between words. The programme of operations is also controlled by numerical codes. The analysis and synthesis of linguistic data necessitates an exact inventory of these data, and this inventory must be made in terms which can be reduced into codes. In this context it is possible to employ the word *metalanguage* in a restricted but active sense, as does Andreev:

1. We call a metalanguage any linear system of signs used for the written designation of the elements in a particular system of ideas and the relations between these elements.
2. The class of metalanguages at the present time comprises mathematics, physics, chemistry, formal genetics, and symbolic logic.
3. A special metalanguage in the symbols of which the facts and relationships of the language systems may be described that are subject to equivalent comparison, needs to be developed for the preparation of algorithms for machine translation.
4. The symbols used in the metalanguage of machine translation are regarded as metalanguage words and grouped in categories analogous to the parts of speech. [28]

Thus the metalanguage envisaged by Andreev is a system of symbols expressing linguistic elements and the relations between these elements, the study of which should make it possible to prepare automatic translation programmes. These symbols would be to automatic translation what H_2O and NaCl are to chemistry. They would be immediately comprehensible to specialists, but of no use to the uninitiated. An international language requiring no translation, they would be immediately transposable into machine codes. In this restricted sense, the idea of a metalanguage resembles that of a strictly closed semantic system of expression such as algebraic or chemical symbols. It is not unlike the "linguist to computer" code of instructions recently devised by M.I.T. research teams under the name of COMIT. Such a metalanguage can never be a substitute for natural languages, but may make it possible to study such languages with a degree of precision not easily attainable by the use of ordinary language, in which meaning is frequently distorted by analogy, metaphor and all the fluctuations of semantics.

Translating-machine metalanguage would be a highly specialized instrument for use by those engaged in facilitating translation; it would help them to define and fix their ideas with greater precision and also to communicate these ideas to the machine in completely unambiguous form; the symbols of this metalanguage would set in motion the machine operations corresponding to linguistic programmes or sub-routines. On this level it is not a question of philosophy, linguistics or logic, but of the creation of an instrument by means of which the human mind would be able to explore the realities of language and at the same time control and discipline the mechanical forces which unconsciously simulate certain functions of the conscious mind. It would be an instrument of communication between certain specialists, an instrument for the control of the machine—a language used as a precision tool, working on language as its object, making possible by its very precision the exploration of this object, at the same time respecting the imprecisions and illogicalities inherent in its nature.

THE LINGUISTIC MOULD OF REPRESENTATION

Applied to linguistic data this instrument should make it possible to reconstitute the *meaning* of the sentence translated, that is to say to enable the reader of language B to participate in the *representation* of the speaker or writer of language A. The question the translator most frequently asks himself is: "What does this *mean?*" and linguistic data are to him inseparable from their meaning. The machine, even more than the human translator, will have to approach meaning, so closely bound up with the representations of writer and reader, by the narrow path of material signs expressing representation within the limits of the expressive systems of the languages employed. These limits are imposed by the choice and meaning of words, by the forms such words may take, by the relationships existing between them: semantics, morphology and syntax all mould and frame thought. Each language is in effect a collective system of expression, a framework or mould largely prefabricated by the social life of a human group. It is rare when the rigid outlines of this mould completely coincide with the individual forms of the representation we are trying to communicate.

Now the moulds for the expression of a thought offered by two different languages are rarely identical. The prefabricated elements of these moulds have neither the same shape nor the same dimensions, except in the case of strictly scientific texts for which scientists have created means of expression not far removed from metalanguage.

Therefore, in order to reconstruct the meaning of a sentence in language A with the semes, grammatical forms and syntactic structures of language B, the machine must identify each of the linguistic elements of the sentence, including the relationship between words, must assign to each of these a code number, which in turn must be able to call for and deliver at output the semes, inflexions and other morphemes, as well as the equivalent syntactic relationships of language B, thereby making a comprehensible sentence in that language. The mesh of different linguistic elements in the sentence in language A, carefully disentangled by the machine, will be replaced in the memory-registers of this machine by a series of codes sorted out into strict order, to be recombined afresh in a different network composed of those elements of language B which reproduce the meaning of the original sentence. How can this be accomplished?

HIEROGLYPHIC CONVERSION

The problem being that of respecting meaning to the maximum extent compatible with the necessity of rendering the sentence pliable to the demands of the machine's codes, the first step is to decompose the sequences of words and their inter-relationships in language A, and to express them by a series of equivalent figures, capable in turn of being transformed later into a meaningful sequence of words in language B. Andreev gives the name "hieroglyphs" to these numerical codes which represent semes, forms and structures. He divides them into three classes: semantic hieroglyphs, formal hieroglyphs, and tectonic hieroglyphs.

> When the input text is coded, numerical symbols are obtained for the ideas contained in lexical units (semantic hieroglyphs), numerical symbols for grammatical morphemes and symbols for link words (formal hieroglyphs), and numerical symbols for word order and syntagmatic relationships between words which are not expressed phonemically (tectonic hieroglyphs). In decoding, corresponding hieroglyphs determine the choice of words, their grammatical formation, and the methods of their combination in the output language, [1]

The main task of the machine—the fundamental linking factor in machine translation—is hieroglyphic conversion. The basic pattern of automatic translation consists of three principal phases: analysis, or coding of information given in the input sentence; conversion, or the substitution of one code for another; synthesis, or the decoding of the converted information into a text in the output language. In bilateral translation programmes of the type A→B, analysis and conversion are carried out simultaneously and are conditioned by the need to arrive at the linguistic framework of language B.

It is, however, possible to envisage other methods, for example that of analysing all the inherent forms and grammatical relationships of language A, entirely disregarding any "output" language B. Language A would be coded in complete isolation, every word being analysed according to all the inherent grammatical forms of that language, whereas in A→B translation analysis is restricted to the *differences* between the characteristics of the two languages.

For instance, to translate into French *her fine clothes,* it is necessary to effect an analysis which will determine the fact that *fine* is plural although without any visible plural sign, so that the French adjective will be made to agree with the noun *habits* when the moment comes for synthesis. If, on the other hand, the machine were translating from English into a language in which adjectives do not agree, such an analysis would be superfluous. Analysis must also determine the number and gender of the word *habits,* so as to obtain correct translation of the possessive adjective *her,* whereas in translating into French it will not be necessary to determine the gender of the possessor, even though this is indicated in English. If the output language declines substantives, further analysis of *clothes* will be required in order to indicate whether the noun which translates it is in the nominative, genitive or any other case. Such analysis is not needed for translation into French.

If the aim of the analysis is to translate from one language into any other language, it is obvious that it must be as complete as possible for each part of speech, whereas for translation of the type A→B, it need only be partial, its extent depending on the degree of parallelism between the structures of the two languages. Given the hypothesis of an analysis for universal application for translation of the type A→X, it will be sufficient, after a complete analysis of language A, to draw up for each pair of languages, conversion tables for the hieroglyphs representing the elements of the analysis. Andreev reminds us that:

> Analysis and synthesis are constant values for every language, and are determined exclusively by the norms of a given language and by the principles of coding. Hieroglyphic conversion, on the other hand, is a variable value, and is a function of both input and output hieroglyphs, [11]

Three main types of difference exists between input and output hieroglyphs:

(1)The suppression of superfluous hieroglyphs: German *die Sprache—language*—coding of the article *die* is superfluous; *her dress, sa robe*—the hieroglyph for the gender of the possessor is superfluous.

(2)Introduction of additional hieroglyphs: Japanese *ani no hon —elder brother's book*—Japanese having no genitive case, an additional hieroglyph denoting this case is required and must be introduced for English; *her dress—sa robe*—the hieroglyph denoting the gender of the object possessed is required and must be added.

(3) Modification of the type of hieroglyph: *to catch cold*—French *s'enrhumer,* Russian *prostudit'sja*—one of the English words, *catch,* is represented by a semantic hieroglyph whereas in French and Russian this must be replaced by a formal hieroglyph denoting the reflexive verb: *to make clear*—*clarifier*—the semantic hieroglyph for *make* must be replaced by a formal hieroglyph of a verbal type.

LINGUISTIC ANALYSIS BY MACHINE

The accounts of Soviet experiments provide the best concrete illustrations of the methods employed to transform words into "numerical equivalents" or codes stored temporarily in the memory-registers of the machine and making it possible to synthetize the sentence in another language. In order to understand these illustrations fully it is essential to be acquainted with the broad outlines of the translation programme drawn up for the B.E.S.M. computer by the Institute of Precise Mechanics and Computer Techniques of the Academy of Sciences of the U.S.S.R.

Figure 2 gives the general layout of the programme: input of English text; analysis, divided into two main phases—vocabulary and parts of speech; synthesis—also divided into vocabulary and parts of speech; output, or printing out of Russian text. Operations take place in descending order. Input is in Baudot telegraphic code.

At this point a brief description of the electronic dictionary here employed must be given. As shown in Figure 2, it has two sections—English and Russian. The first part contains, in numerical code, the English words accompanied by:

(a) Certain permanent information concerning each word, including an order number indicating its place in the English section of the dictionary.

(b) Where applicable, the order number of the corresponding Russian word, making it possible to obtain certain permanent information concerning this word.

(c) If the Russian word is not given, an instruction referring back to the polysemantic dictionary;

(d) Appropriate instructions concerning the next sub-routine to be performed.

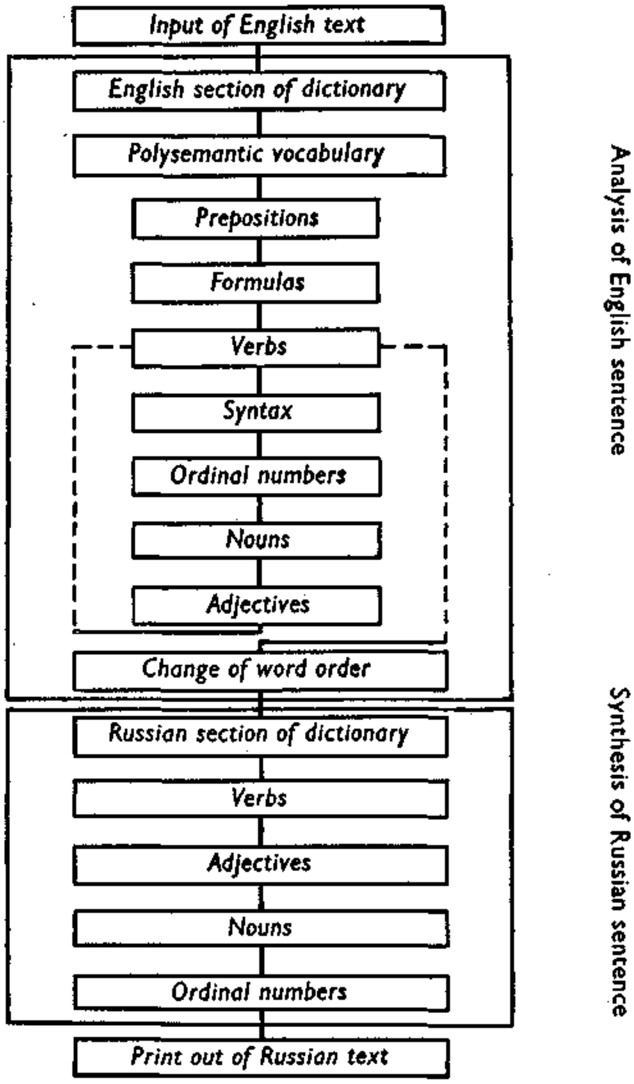


Fig. 2. Block-diagram of automatic translation programme from English to Russian. After Korolev, Rasoumovski and Zelenkevitch, *Les experiments* (sic) *de traduction automatique de l'anglais en russe a l'aide de la calculatrice BESM*, USSR Academy of Sciences, Moscow, 1956 (in French).

The Russian section of the dictionary contains the numerical codes representing the letters which compose the Russian words, and the permanent information concerning these words. It enables the machine to construct Russian sentences at output, by reassociating the stems of these words with the endings in accordance with the indications recorded in the memory register during the analysis of the English sentence.

English words having more than one meaning are placed in a polysemantic or supplementary dictionary which makes it possible to choose, from among several meanings of the same word, the correct Russian equivalent for the given context. When the machine looks up a polysemantic word in the English section of the dictionary, it finds not a Russian order number but a code instructing it to look for this word in the polysemantic dictionary, where it will receive supplementary instructions setting in motion the particular sub-routine for deciding the meaning of this word (see below, sub-routine to determine the meaning of *of*).

If the spelling of an input word corresponds exactly to that of a word in the dictionary the word is identified and the programme continues immediately in accordance with the dictionary indications. If, on the contrary, the word is not in the dictionary (e.g. the word *walked)* the machine immediately embarks on the sub-routine for the reduction of inflected endings (see Figure 3) which, in the case of the English perfect tense, makes it possible at the fourth attempt to detach the ending *-ed* and subsequently to look up and identify the "base" *walk* in the dictionary.

We must now return to Figure 2 and the analysis of the English sentence. This analysis consists of a series of sub-routines, all designated by the names of parts of speech, with two exceptions— the sub-routines for *formulae* and *syntax.* This programme was designed for the translation of mathematical texts, and mathematical formulae (in which Russian M.T. scholars include foreign words and some proper names) require no translation. The name for the "syntax" routine is self-explanatory.

For the various parts of speech, a special analytical sub-routine aims at identifying and inscribing in the machine's memory-register, in coded form, all the information necessary for the construction of a grammatically correct Russian sentence equivalent in meaning to the English sentence. When a word is identified in the dictionary as being a preposition, the sub-routine for prepositions is put into action, and the machine proceeds to a series of check-ups concerning the context of the word; such check-ups are to determine whether the word is to be translated into Russian or not, what case is governed by the Russian preposition which translates it, or, if it is not to be translated, what case the following word should be, etc.

Similarly, the sub-routine “English nouns” submits the English noun or pronoun on which the machine is working, once it has been identified as a substantive or its equivalent, to a series of check-ups to determine its grammatical role, its case, gender, number, etc. These check-ups are made by the method of dichotomy: that is to say by asking a series of questions which can receive either a negative or an affirmative answer. If the reply is affirmative, the machine classifies the information received in its memory-register. If it is negative it continues its search in a predetermined order. In the memory-register two “cells” are allocated to each word: in the first cell is inscribed on the right the order number of the English word, then, in fixed positions from left to right, the affirmative or negative indications received in reply to each of the questions applicable to this word. The sub-routines are performed in a pre-established order enabling all the divisions of the cell to be filled in. The number of divisions varies with the parts of speech concerned: two for adverbs and prepositions, four for conjunctions, eight for cardinal numbers, seventeen for nouns, eighteen for adjectives and verbs. These divisions are always followed by a four-figure division for the order number of the word. The second cell contains only the order number of the equivalent Russian word in the Russian section of the dictionary.

It will be observed that the sub-routine for ordinal numbers—*"numerical adjectives"*—is justified not by any peculiarities in their behaviour in English, but by the fact that in Russian they are declined. Let us observe, too, the dotted line that links “verbs” and “adjectives”: the Russian verb has numerous adjectival forms which are declinable. Their behaviour is therefore both verbal and adjectival, with declension and conjugation, and it is therefore necessary to return to the verbal sub-routine after analysis of certain adjectival forms such as English participles.

Once the analysis is complete, each English word in the sentence appears in the memory register in the form of an order number, accompanied by all the coded information defining its grammatical role and all the information required to decline or conjugate the corresponding Russian word. Panov calls this the *numerical equivalent of the English word.* It remains only to apply the sub-routine "change of word-order", according to the indications contained in the divisions of the first cell.

SOME EXAMPLES OF SUB-ROUTINES

A few examples will help us to visualize in concrete fashion some of these sub-routines. The first example [23] illustrates a case of multiple meaning: the determination of the exact meaning in Russian of the English preposition *of.* In the table given below the notation 1.(3,2) signifies that if the result of the first operation is affirmative operation (3) should be performed; if the result is negative, then the machine should perform operation (2), and so on. The notation 3.(0, 0) signifies that the sub-routine is terminated and that the result should be recorded in the memory-register.

1.(3, 2) Check preceding word for *is, are, was, were, be.*
2.(3, 4) Check following word for formula, or *course.*
3.(0, 0) This particle is not translated.
4.(5, 6) Check preceding word for *idea* or *discussion.*
5.(0, 0) Prepositional case: Russian translation *o.*
6. (7, 8) Check preceding word for *true* or *productive.*
7.(0, 0) Russian preposition *dlja* followed by genitive.
8.(9, 10) Check preceding words for *fall short* or *in place.*
9.(0, 0) Preposition not translated. Following noun in genitive case.
10.(11, 12) Check preceding word for *out.*
11.(0,0) Genitive case: translation *iz.*
12.(13, 14) Check preceding word for *incapable.*
13.(0, 0) Preposition *k* plus dative case.
14.(15, 16) Check following word for *necessity.*
15.(0,0) Dative case: translation *po.*
16.(17, 17) Genitive case.
17.(18, 20) Check preceding word for noun.
18.(19, 20) Check following word for noun, cardinal number, or formula not followed by a noun.

19.(0, 0) Preposition not translated. Following word in genitive case.
20.(21, 22) Check the preceding word for *consist, each, one, some* and the following word for *all* or *them.*
21.(0, 0) Translation *iz.*
22.(0, 0) Translation *ot.*

Close study of this example will show that most possible translations of "*of*" are included, beginning with those occurring least frequently and ending with those most often encountered.

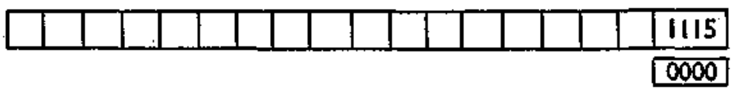
Grammatical analysis. Another series of examples will show how the figures representing grammatical characteristics which will, at the appropriate moment, control the synthesis of the Russian sentence, are inscribed in the memory register. Panov, in his brochure *Automatic Translation* [25], demonstrates his method by means of the following English sentence:

> This is true certainly of the vast category of problems associated with force and motion.

We shall here limit ourselves to examining his illustrations for the words *this, certainly, of* (twice) and *associated,* completing them in certain details with the aid of Muhin's brochure. [23]

This

The machine looks up and finds in the dictionary the English order number—1115—but fails to find a corresponding order number for a Russian word. The absence of such an order number refers it to the supplementary dictionary for polysemantic words. The information entered in the cell now reads as follows:



The polysemantic dictionary sets in motion the sub-routine for determining the Russian meaning of *this* which will make it possible to fill in the first cell of the memory register by five figures signifying: 1: noun; 1: singular; 3: neuter; 1: nominative; 0: hard stem.

1			1		3					1	0			1115

Having now been identified as a noun for the requirements of the translation programme, the word is submitted to the check-ups prescribed by the sub-routine “English nouns”. As a result the Russian order number of the word *eto* is entered in the second cell and the first cell is completed as follows (reading from left to right):*1: noun; first 0: is declined like an adjective;† 1: invariable; 1: singular: 1: nominative; 3: neuter; 1: there is an indication of number; 1: subject; 0: hard stem (in Russian); the last 0 signifies: absence of an instruction to "omit word".

1	0	0	1	1	0	1	3	0	0	1	0	0	1	0	0	0	1115
																	6327

Certainly

The dictionary immediately provides the order number for the Russian word *bezuslovno*—2257—together with the indications 5: adverb; 1: interpolated word; English order number 0132. The Russian order number being given, there is no need to consult the supplementary dictionary. The cells are filled in as follows:

5	1	0132
		2257

Of (the vast category)

The dictionary indicates: preposition, English order number 0472: it orders the sub-routine “English prepositions”: the cell reads as follows:

6		0472

the 6 indicating “preposition”. In the course of executing the sub-routine for this preposition, which involves determining which one of its multiple meanings is required, the machine finds

*Zeros in sub-divisions 3, 6, 9, 10, 13, 15, 16, of the cell are not relevant to the case under consideration—although of course each sub-division must be filled, either by a zero or by another figure.

† This means in fact that this ‘noun’ *eto* is a pronoun, and the following sub-division indicates that *eto* is invariable.

(operations 6 and 7 of the example quoted above) the translation *dlja* followed by the genitive, and enters in the second cell the Russian order number for *dlja,* 5046. The first cell is now filled in as follows:

6	2	0472
		5046

the 2 signifying "governs the genitive"

Of (problems)

The machine operates exactly as for the preceding example and in executing the sub-routine for prepositions receives the answer "is not translated, governs the genitive" (see above, operations 17, 18, 19). The absence of a Russian order number in the second cell means that the word *of* will not be translated at output, but the inscription of a 2 in the appropriate division of the first cell will ensure that the following word (the translation of *problems)* is declined in the genitive case. The first cell will thus be filled in exactly as in the case of the preceding *of,* and the Russian cell will be empty, reading as shown below:

6	2	0472
		0000

Associated

The dictionary check up for this word revealing no equivalent, the programme for the separation of inflected endings is put into action (see Figure 3); when the ending *-ed* has been eliminated, the word *associat-* is found, accompanied by the following information: 2: verb; 1: 1st conjugation; 4: governs the accusative; 0: imperfective aspect; 3: has a desinence -ed; English order number 0085; the fact that there exists a Russian order number (2140) refers the machine to the sub-routine "English verbs". The cell is now partially completed thus:

2		1					4		0								3	0085
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	
																		2140

The machine executes the sub-routine “English verbs” and finds the information shown below in its numerical form in the 18 sub-divisions of the cell and explained step by step. The figures in brackets are the numbers of the sub-divisions, which have been added in order to facilitate reference.

3	1	1	0	0	0	0	2	2	0	1	0	3	0	0	0	0	1	0085
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	
																		2140

Sub-division number

(1) 3: adjective (the participle being the adjectival form of the verb, the 2 originally recorded in this division is now replaced by a 3);
(2) 1: soft stem,
(3) 1: 1st conjugation (we *are* dealing with a verb and this information is needed for synthesis of the participle);
(4) 0: stem ending neither in a sibilant nor a guttural;
(5) 0: variable word;
(6) 0: plural;
(7) 0: not a predicate;
(8) 2: genitive;
(9) 2: feminine;
(10) 0: designates an inanimate object;
(11) 1: takes the shortened form;
(12) 0: has no indication of number;
(13) 3: participle;
(14) 0: past;
(15) 0: not subject;
(16) 0: not governing a case;
(17) 0: no indication “omit this word”;
(18) 1: the English word has ending *-ed.*

ANALYSIS AS PRE-SYNTHESIS

These examples demonstrate how analysis and hieroglyphic conversion are combined in the programmes illustrated: we are dealing with bilateral programmes of the type A→B, and not with a universal programme of the type A→M→B. Here, analysis is pre-synthesis—the study of the English sentence being directed towards and conditioned by the needs of the Russian synthesis. If the machine records a code meaning “feminine gender” after the word *associated,* it is only because the Russian word corresponding to *problems* is feminine.

It is perfectly normal and legitimate thus to undertake a thorough check-up on the role of the English words in the sentence solely with a view to synthesis into Russian. But it will be observed that to translate into French the word group *the vast category of problems* would require a quite different and much simpler analysis. The Russian translation omits *the* and *of,* French would translate them. Russian declines *vast, category, problems.* In French these words present problems of number and gender only. Analysis for translation into French would be shorter and simpler than for translation into Russian. The same is true of certain polysemantic problems: in both instances the word “*of*” would be translated by *de* in French.

TOWARDS A MULTILATERAL PROGRAMME?

It is, however, possible to see how the Russian method illustrated above, although bilateral, is potentially capable of generalization. When the memory cell for each word of a sentence has been filled in, the analysis being complete and the synthesis not yet begun, the machine is in the same state as a translator who has looked up all his words and analysed all their inter-relationships, but has not yet begun to write. His translation is “in his head”. The machine has “in its memory” numerical hieroglyphs representing semes, grammatical forms, syntactic structures—which it can, at a given signal, transcribe into Russian words correctly ordered, conjugated and declined.

Is it possible that instead of Russian words (language B) it could align French words (language C) corresponding to the English sentence (language A) and to the hieroglyphs representing it?

The machine can certainly do this if the number of categories according to which the analysis has been made in view of language B is superior or equal to the number of categories necessary for language C, or conceivably if it brings into play a supplementary programme of analysis B→C when language C requires a more complex analysis than that which was required for language B.

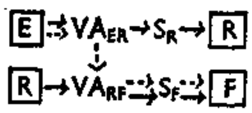
It is equally clear that a universal programme must include in its analysis all conceivable grammatical categories for each part of speech and all their combinations in order to provide all the necessary indications for synthesis into any target language. Only practical research spread over a number of languages can make it possible to draw up such a programme, and this research must at first be of a bilateral type, i.e. language A into language B.

The Academy of Sciences of the U.S.S.R. sees in its methods of analysis and synthesis a practical and almost immediately realizable solution to the problem of an intermediary language. Panov has illustrated this belief in the following formula. If E stands for English, R for Russian, F for French, V for vocabulary, A for analysis and S for synthesis, the phases of the programme E→R can be represented as follows:

If the double operation of English-Russian vocabulary and analysis is summed up by the symbol VAER, it will be seen that, once this stage is passed, the Russian equivalents of words are available, ready for synthesis.

"It is evident", writes Panov, "that when once VA is accomplished, we are fully provided with Russian equivalents, with all their grammatical indications attached. Thus we can easily pass on to, say, a Russian-French dictionary and get the corresponding French equivalents together with their necessary grammatical indications. Thus using only the French synthesis programme we can obtain a French sentence automatically translated from English." [26]

It is, in effect, possible, by using a synthesis programme S_F and a Russian-French dictionary, to obtain the following formula, in which the continuous lines represent the normal programme and the dotted lines the proposed variations:



Russian is, of course, the interlanguage proposed in this solution, or rather that kind of pre-Russian constituted by analysis of E→R in which Russian grammatical categories are superposed on or substituted for those of English. It is clear that Russian, being a synthetic language and rich in grammatical categories, is thus particularly well placed to be used as the basis for a machine-interlanguage. Since in Russian the indications of grammatical categories are attached to the words themselves instead of being concealed in the intricacies of word order, Russian is better able to respect the individual concrete character of the original sentence than is a language less rich in inflexions and in which word order, which is necessarily rigid, plays an important part in communicating the meaning of the sentence.

The method of multi-lingual translation proposed by Panov is ingenious but not without its drawbacks. It is based on an analysis of English, with Russian as the goal and as the tool. The figures are recorded in the memory cells according to a system based upon the Russian language. Since he who can do more can do less, it is possible that this system may give satisfactory results for translation into languages less richly inflected than Russian. But this method does not answer the objections raised by Booth. Is not a programme E→F preferable, if only for reasons of economy, to a programme E→R→F? Such a programme, while perhaps of some value in Moscow, London or New York, would certainly not be suitable, for instance, for a French organization engaged in the production of scientific translations into French from a number of foreign languages. For translation from English, Romance languages, etc., such a team could utilize analysis programmes much simpler than those designed by the Russians, though based on the same principles. This simplicity would be due to the greater degree of resemblance between the structures of the languages involved.

PRIORITY OF BILATERAL PROGRAMMES

The Russian method makes it possible to have only one single synthesis programme for any target language. Thus a national workshop translating into the native language from a number of foreign languages has everything to gain by drawing up its own synthesis programme and by preparing for each source language an analysis programme in conformity with the demands of the target language.

For theoretical as much as for practical reasons, bilateral programmes are at present the vital ones. The sum of knowledge required for the eventual establishment of universal programmes can be acquired only by comparison of numerous bilateral programmes, which should be worked out as the first priority in all countries interested in machine translation for practical purposes. The execution of translation work at high speeds and without waste of precious time will depend on the economy of the programmes, that is to say on the degree to which they can avoid all check-up routines having no immediate utility for the translation in process. Even if it takes but a thousandth or even a millionth of a second to fill a sub-division of a memory cell, economies of a thousandth or a millionth of a second will be of great value in programme making, as they will occur an infinitely great number of times in each translation.

One of the great merits of the Russian system is that it makes possible the use of whole sections of bilateral programmes for work on different languages. With progress in self-programming it is possible to look forward to a day when the machine will itself choose from among several programmes those applicable to a given language, its choice being governed by considerations of maximum economy. At this point Andreev's metalanguage would tend to become one and the same thing, at least in this context, as a recorded set of instructions enabling a machine to choose, from among a universal but empirically elaborated programme of translation, those elements required for the translation of language A into language B and vice versa, and to perform this translation with the greatest possible measure of economy. But much time-consuming practical work is essential before this point is reached.

CHAPTER V

Syntax and Morphology

DURING the early days of research the priority of lexis over morphology in preparing the way for machine translation was taken for granted. In drawing up the first automatic dictionary, Booth and Richens gave scarcely a thought to grammar. Its true importance became evident only as partial, word-for-word translation was abandoned little by little in favour of an attempt to produce genuine, fully automatic translation. In the meantime the work of preparing programmes for the machine clearly revealed certain weaknesses in traditional grammars and opened up new horizons as to possible fresh classifications of linguistic data.

Reifler's work in Seattle, that of the Wundheilers at the Illinois Institute of Technology, of Bar Hillel, Chomsky and Victor Yngve, Oswald and Fletcher's work on German grammar, expanded later by Booth and Brandwood in London, Brandwood's further work on French—the accumulated experience and conclusions of all this research certainly contributed considerably to the subsequent rapid progress made by Soviet linguists and technicians. The role of morphology and of syntax was henceforward clearly defined. Automatic analysis of the function of the word in the sentence, illustrated in the preceding chapter by Soviet examples, was rendered possible by the previous research of these pioneers.

IMPORTANCE AND LIMITS OF GRAMMATICAL PROBLEMS

As we have already seen, the somewhat naive optimism of the early stages as to the usefulness of providing rough word-for-word translations soon came up against the problem of inflected endings, and later against the necessity of determining the grammatical values of uninflected words. It soon became clear that a sequence of semes is insufficient to communicate the meaning of a sentence, even to a specialist in the subject matter concerned.

Translation was impossible without prior knowledge of the grammar of the original language. One major reason for this was the high proportion of polysemantic words.

Setting to work on this problem, Yngve [17] took as his starting point the improvement of word-for-word translation and evolved a series of basic principles. The information necessary for the solution of the problem of multiple meaning, he observes, resides in the context, that is within the sentence itself. Basing himself on Zipf's work on word frequency, he noted that words very frequently used are those that have the most meanings. The fifty most frequently used words account for about one-half of the running words of a text, so that it is clear that a solution to the the problem of multiple-meaning for the fifty most frequently occurring words would go more than half-way towards solving all problems of multiple meaning. These common polysemantic words proved to be grammatical tools—or "cement words"—articles, prepositions, conjunctions, auxiliary verbs, pronouns, etc.—the very words which constitute the grammatical structure of language in which the nouns, verbs, adjectives, adverbs, etc., are contained. Each sentence being almost separate, both grammatically and syntactically, from its neighbours, a sentence would probably be a suitable basis for translation and it would rarely be necessary to go further afield to find a solution for the problems of multiple meaning involved in the fifty most frequently used words.

To lend support to arguments based on Zipf's law, Yngve produced evidence likely to convince unbelievers by recounting in detail the results of an experiment he had himself conducted. A partial translation of German was prepared, taking as text 750 running words of a review of an American work on mathematics. The 250-word vocabulary of this text was put on cards and Yngve, without prior knowledge of the text, translated each card in alphabetical order. The mathematical vocabulary was correctly translated but words like *der* and *sein* proved to be untranslatable out of their context. They were therefore left in the original German, and German word-order and flexional endings were also left unaltered. The "translation" was typed out, the English words in capitals and the remaining German elements in lower case letters. Here is part of the result:

"Die CONVINCINGe CRITIQUE des CLASSICALen IDEA-OF-PROBABILITY IS eine der REMARKABLEen WORKS des AUTHORs. Er HAS BOTHen LAWe der GREATen NUMBERen ein DOUBLEes TO SHOWen: (1) wie sie IN seinem SYSTEM TO INTERPRETen ARE; (2) THAT sie THROUGH THISe INTERPRETATION NOT den CHARACTER von NOT-TRIVIALen DEMONSTRABLE PROPOSITIONen LOSEen. CORRESPONDS der EMPLOYEDen TROUBLE? I AM NOT SAFE, THAT es dem AUTHOR SUCCEEDED IS, den FIRSTen POINT so IN CLEARNESS TO SETen, THAT ALSO der UNEDUCATED READER WITH dem DESIRABLEen DEGREE OF EXACTNESS INFORMS wird . . ."

The full text of this partial translation was then put before two categories of reader: those who knew no German, who were able to understand the subject matter only, without grasping the meaning, and those who knew some German grammar and who, once they had recovered from their amusement, demonstrated that they had understood rapidly and well.

The next principle enunciated by Yngve was therefore: "Concentrate on the grammatical problems since they account for the majority of multiple meaning problems, and the specialized field glossary can cope with most of the rest."

This experiment also demonstrated another point—namely, the importance of word order. Since the problems of multiple meaning are bound up with problems of syntax, the context of language B can help us to understand the meaning of words in this language only if the word order of language B is respected in translation from language A into language B. Morphology and syntax can generally provide the solution to those problems of multiple meaning which occur most frequently.

This brief account illustrates two facts with which all subsequent work on automatic translation has had to reckon. Morphological and syntactic problems are paramount and no genuine translation is possible until those problems have been solved. They are, however, relatively restricted in scope, and once the solutions have been found for a given pair of languages A→B, they will be applicable to all translations of A into B.

An experiment recounted by Bel'skaja [5] strongly confirms the fact that morphological and syntactical problems are at once of primary significance and strictly limited scope. In an important article in *Research* she describes the first English-Russian dictionary used by the Academy of Sciences of the U.S.S.R. for its experiments. This dictionary comprised about 5,000 words, almost equally divided into English and Russian. Since it was designed for work on Russian mathematical texts, the vocabulary was, of course, highly specialized. "As to the grammar part of the translation programme", writes Bel'skaja, "it has very little, if at all, been affected by the fact that a very limited field, that of mathematics, had been chosen for machine translation. *Indeed, the grammatical programme has proved to be universally applicable.*"

In support of her contention, the Soviet linguist quotes some remarkable experiments made in order to discover whether the same grammatical programme could be applied to a text as far removed from mathematics as, say, an article from *The Times,* or a passage from Dickens. "The experiments have proved the success of our ideas on the possibility of having a universal grammatical programme for the machine translation of any two languages; in the vocabulary field a series of specialized dictionaries, covering different fields of human activities, are unavoidable."

The only thing with which the machine can be reproached in this experiment is that it was obliged to leave in English the words for which no translation had been entered in its dictionary—but in a sentence such as the following only one English word (in italics) remained in the Russian text:

"It made a great impression on me, and I remembered it a long time afterwards, as I shall have occasion to *narrate,* when the time comes."

"Eto proizvelo bol'šoje vpečatlenije na menja, i ja pomnil eto potom, dolgoe vremja, kak ja budu imet' slučaj *rasskazat',* kogda pridet vremja." *(David Copperfield,* Chap. XVI.)

Here we have proof that grammatical analysis of the sentence is valid for all sentences, provided they respect the grammar and above all the syntax of the language in question. A grammatical programme established for two languages A→B can therefore be used for all translations from A to B. It must, however, include all constructions normally employed in that language. Machine-translation linguists will have to pay more attention to morphology in languages which are analytical and richly inflected, whereas in synthetic, poorly inflected languages the problems of syntax will be paramount.

MORPHOLOGY AND THE MACHINE

Figure 3 shows how the problem of exploring word forms has been solved in a language as poor in inflexions as English. English has only six flexional endings: -*'s* or *–s'*, -*s* for the plural or 3rd person singular of the present tense, the verbal-endings -*ing* and -*ed*, -*er* for the comparative, -*est* for the superlative and -*th* for ordinal numbers. The Russians added a false inflexion, essential for the operation of their dictionary: -*e* as in to lov*e*, to cloth*e*, etc., and in adjectives like tru*e*. These words figure in the dictionary as *lov-*, *cloth-*, *tru-*, thus permitting the identification of *lov-es*, *lov-ed*, *lov-ing*, *tru-er*, etc.

Similarly in the French translation programme drawn up by Mel'čuk and Kulagina on the basis of mathematical works by Paul Appel and Emile Borel, stems or bases* and flexional endings have been entered separately in the dictionary, the determining criteria being empiric in nature and not historical. The base of *travail* is, for instance, entered as *trava-* because it appears in the two forms trava-il and trava-ux; *parler* is given as *parl-*, *finir as fini-*, and so on. The common verbal endings have been grouped and classified as in Brandwood's work on French [7] in ending tables which make it possible to analyse the word grammatically as soon as the base has been identified.

Thus both for source and for target languages, for each variable part of speech—nouns, verbs, adjectives, pronouns, ordinal numbers, etc.—the machine will have an ending table making it possible both to separate the bases after input and to identify them, and to conjugate and decline the words at output by adding the endings to the bases. For the separation of endings, the machine will always follow the plan laid out in Figure 3—the only difference being that Russian or German will require more sub-routines than English, in which the maximum number of attempts was seven until Bel'skaja added various refinements

* "Base" is a more correct name than "stem" for the graphically invariant part of an inflected word.

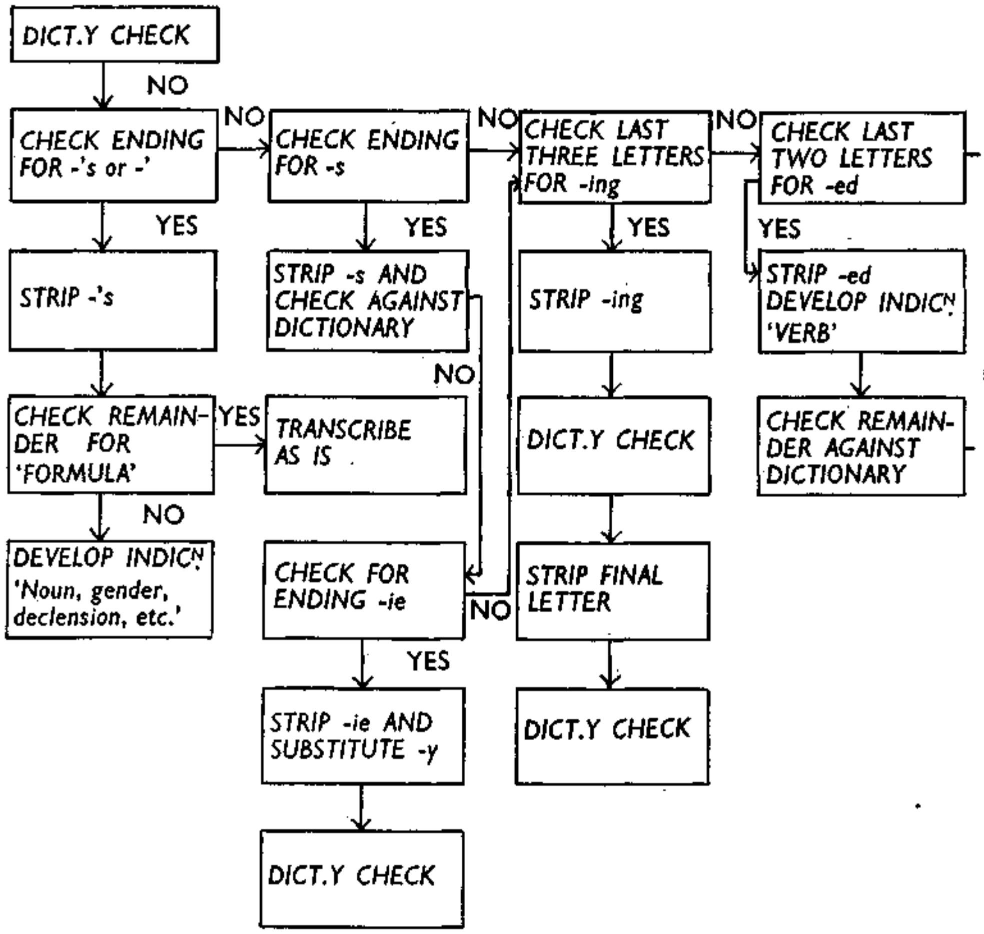
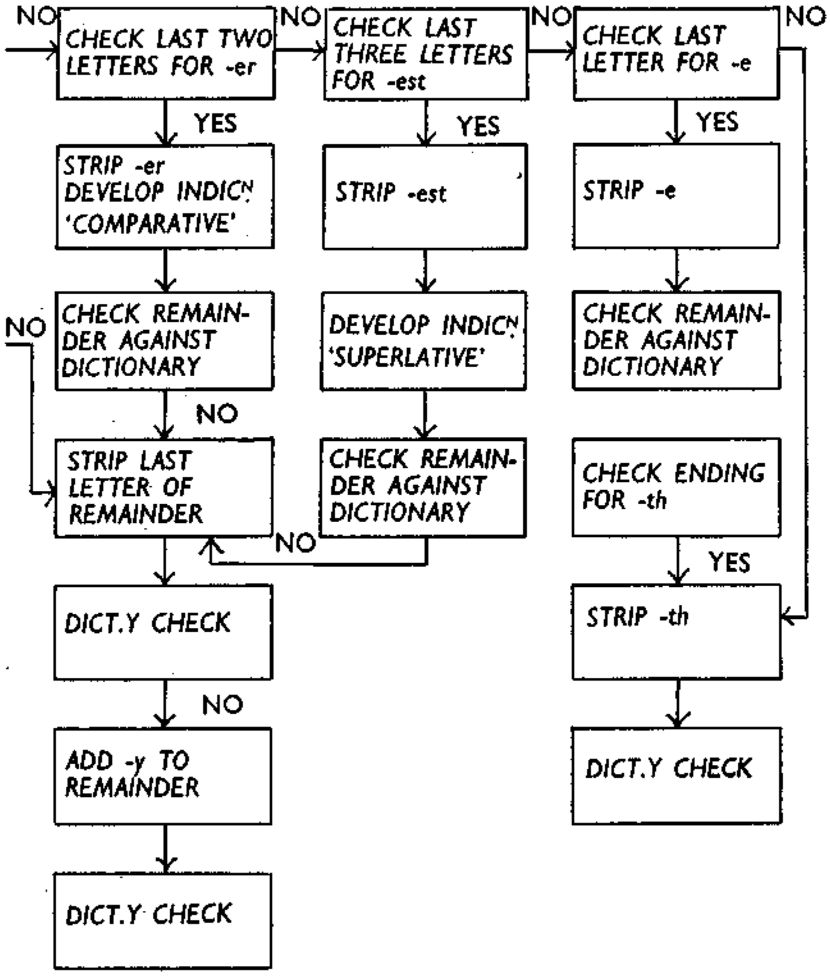


Fig. 3. Routine for stripping English word-endings and dictionary check of remainders. After I. C. MUHIN, *Experiments in Machine Translation with the BESM Electronic Computer*, USSR Academy of Sciences, Moscow, 1956 (*in Russian*).



ABBREVIATIONS
DICT.Y = Dictionary, INDICN = INDICATION

making it possible to trace the bases of Latin words currently used in English, as well as plurals like *busmen, carmen,* etc. For the synthesis, when endings are added to a base without its being modified, as in *aimer-ai, -ons,* etc., a single base will be entered in the dictionary. When the base is modified by certain endings, it is preferable to enter both forms in the dictionary; alternatively, when a base is regularly modified (as in the case of Russian words ending in sibilants or gutturals) the machine can be instructed to alter certain consonants before certain endings. Here again the only criterion is programme economy.

In all poorly inflected languages—and even in those with a number of inflexions—the endings do not always provide information on the grammatical role of the word in the sentence. As Mel'čuk and Kulagina [18] have observed, French contains forms which can be either verb or noun: *la forme, il forme, le fait, il fait, la limite, je limite,* etc. The ending fails to provide the required clue, just as it fails to distinguish between *il limite* and *je limite.* Here we must turn for a solution to the form classes established by the structural linguists Bloomfield, Harris and Fries. The question is to determine objectively to which form class the word *limite* belongs in any given sentence. It is either a singular noun, or a 3rd or 1st person singular verb (this the Russians call *4th person* until a sub-routine has been able to determine whether it is 3rd or 1st).

When a word, such as *limite,* can be either verb or substantive (VS) the following possibilities exist:

(A) VS preceded by a determinant (article, demonstrative or possessive adjective) other than *le, la, en, is a substantive.*

(B) VS preceded by *le, la, en:*

 (1) if the construction is: Nominative or accusative noun or nominative pronoun, + *le, la, en,* +VS,

  (a) in the absence of a comma or a co-ordinating conjunction before *le, la, en,* VS *is a verb.*

  (b) if there is a comma or a co-ordinating conjunction in this construction,

   (i) VS *is a verb* if there is no verb between it and the end of the sentence or between it and a conjunction,

   (ii) VS *is a substantive* if it is followed by a verb.

(2) if VS preceded by *le, la, en* is not part of a construction of this type then it *is a substantive.*

(C) In all other cases VS *is a verb.*

This is only a preliminary analysis, making no claim to completeness or finality, but illustrating a method inspired by structural linguistics. While Reifler divides German words into "form classes" and records them in separate memories, the Russians of the Steklov Mathematical Institute enter in the dictionary a distinctive numerical indication for each class, this number being in fact an instruction code which refers the machine to the appropriate sub-routine.

STRUCTURAL ANALYSIS

By analysing typical structures of whole sentences, or parts of sentences, it can be made possible for the machine to translate uninflected or partially inflected words. Syntax takes over when morphology offers no solution. The machine then analyses the positions of words in relation to one another. Here are some examples from Mel'čuk and Kulagina.

"*Pas, point,* are negative particles if they come immediately after the verb, or are separated from it only by an adverb, and in constructions of the type *ne+pas+infinitive.* In all other cases these words are substantives.

"*Ensemble* after a determinant or a preposition (from which it may be separated by adjectives, adverbs and co-ordinating conjunctions) is a substantive; otherwise it is an adverb."

The formal rule is here incomplete. *Ensemble* is not an adverb in "*Ensemble* de premier ordre, les Petits Chanteurs à la Croix de Bois ont. . . ." The rule requires completion for cases where *ensemble,* not preceded by a determinant, is followed by an adjective, or else a general rule on appositions must modify this rule. Nevertheless, we have here excellent examples of rules which can be embodied in dichotomic sub-routines.

A more complete example is that of the English noun (and pronoun) analysis, as practised in the Panov-Bel'skaja translation programme. As in the sub-routine quoted in the previous chapter, 1.(2, 7) means: "Perform operation 1. If the reply is affirmative, proceed to 2. If it is negative, proceed to 7."

English Nouns

1.(2, 7) Check given word for *us.*
2.(3, 5) Check following word for noun.
3.(0, 0) Produce sign of dative case.
5.(6, 13) Check immediately preceding word for *let.*
6.(0, 0) Produce sign of nominative case.
7.(8, 13) Check given word for *it.*
8.(13, 10) Check *it* for presence of sign of gender.
10.(0, 0) Take gender from nearest preceding subject.
13.(14, 15) Check for presence of sign of singular or plural number.
14.(0, 21) Check for presence of any sign of case.
15.(16, 19) Check for ending -*s.*
16.(17, 17) Produce sign of plural number.
17.(18, 14) Check preceding word for formula without the sign=.
18.(0, 0) Produce sign of genitive case.
19.(16, 20) Check preceding word for *much.*
20.(14, 14) Produce sign of singular number.
21.(22, 23) Check preceding word for *let.*
22.(0, 0) Produce sign of nominative case and subject.
23.(24, 28) Check immediately preceding word for sign of similar conjunction.
24.(28, 25) Check word immediately preceding and following similar conjunction for adjective.
25.(26, 27) Check all words for same word as the given word.
26.(0, 0) Take case from noun found.
27.(0, 0) Take case from nearest preceding noun.
28.(18, 29) Check for ending -s.

We see here how, in order to determine the case of the noun or pronoun, the machine performs a series of explorations of immediate context of the word in question and of the structure of the sentence. The sub-routine is, in fact, based on Yngve's recommendation: "to make the needed information that is implicit in the context explicit at each word position in the sentence."

CLASSIFICATION AND COMPARISON OF STRUCTURES

In richly inflected languages where word order is relatively free, morphological endings provide the information needed; in languages like English or Chinese where the sequence of words is strictly ordered, analysis must depend mainly on structure and word order. Thus programmes or sub-routines appropriate to the types of the two languages with which we are dealing will have to be prepared for all the basic structures of each and for the conversion of the structures of language A into those of language B.

For the preparation of such programmes, the work of Jespersen, Bloomfield and Fries has proved most useful. From our point of view, the great merit of their researches, undertaken long before there was any question of machine translation, was that they examined grammatical structure without reference to meaning, and sought to define structures independently of meaning, because, as Fries pointed out, meaning provides no means of identifying and distinguishing structures. This applies also to the machine which, when faced with the sentence "*the maid gave the cat meat*" is incapable of identifying the grammatical role of *cat* or *meat* except by reference to structure, whereas any schoolboy learning English will guess their respective roles by the meaning of the sentence.

The work of Fries, which was largely directed towards the teaching of English to foreigners, leads even more surely than Jespersen's *Analytic Syntax* to the expression of structures in algebraic formulae.

English words have been grouped into four classes defined according to their role in the structure of the sentence. Although these form classes do not exactly coincide with the four main parts of speech, it can be said, for the sake of brevity, that Class (1) consists of most nouns, (2) of the majority of verbs, (3) of most adjectives and (4) of almost all adverbs. Fries has also identified fifteen groups of auxiliary or functional words designated by letters A to O, and comprising in all 154 words. Complete sentences can be expressed algebraically by means of these nineteen symbols, to which must be added + for the plural, — for the singular, -d for the preterite, -ing for the present participle, etc.

The sequence of symbols $D\ 1_{\pm}\ 2_{\pm}\text{–d}\ 4$, for example represents sentences of the type: *The pupils ran out. The ships sailed away.*

Soviet linguists have made a thorough study of this system and have classified Russian words into seventeen groups which they have related to the English language classes and groups. Mološnaja [19] has shown that structure-formulae can be grouped according to types or models.

Elementary units of English words, associated in structural patterns, having been classified, these structures were translated into Russian as simply and directly as possible. The Russian constructions were then analysed, reduced to formulae and compared with the equivalent English constructions, whenever possible by two-member word combinations. When two English elements failed to coincide with two Russian elements, additional symbols were introduced as required. For instance in the following English absolute participle construction, the symbol J (subordinating conjunction) has been added to express the difference in structure:

The rain having ruined my hat, I had to buy a new one.
Tak kak dožd' isportil moju šljapu, ja dolžen byl kupit' novoju.

English	*Russian*
1 2-ing 12	J 1 2 1 2

These are cases where an English construction may have several Russian equivalents:

(He) looks pale: 2±3	(On) *vygljadit blednym:* 2±3
(the stone) lay deep (in the water): 2±3	*(kamen') ležal gluboko (v vode):* 2 ± 4^x
a group of children: 1 F 1	gruppa detej: 1 1
a book with pictures: 1 F 1	kniga s kartinkami: 1 F 1

Structures being treated as entities like words, one can thus make a dictionary of structures and identify cases of polysemantic or homonymic structures requiring the formulation of special rules to identify their target language equivalents, these special rules being treated as sub-routines just like those for resolving grammatical homonymy.

Complex structures can be simplified by reduction. For instance

The old man

$\underline{D \quad 3 \quad 1}$ becomes $D\underbrace{\lfloor 3 \quad 1 \rfloor}_{1}$ since the adjective plus the noun are reducible to a nominal unit; this is very useful since in Russian these three words can be translated by a single Class 1 word, and in French by two (determinant and noun).

Thus it is easy to see how, once all possible English syntagmas have been classified and inventoried and the same processes completed for a second language, for example Russian, it will be possible automatically to identify each structure in a complex sentence, to reduce it to a simple formula, which is then converted into a language B formula and later expanded according to the rules peculiar to that language, After reduction, conversion and expansion of the formulae, it will remain only to "unroll" the Russian sentence by means of the semes stored in the machine's memory and the morphological indications which were entered in the memory when the sentence was analysed.

This method of syntactic analysis requires further intricate research and much elaboration of detail. It has been advanced by Mološnaja as a working hypothesis which the machine must verify by application to a large corpus of text. It would appear to be capable of providing a solution to the problems of automatic analysis by the machine of the syntactic elements constituting complex sentences.

It might make it possible to complete sentence analysis by separating temporarily, where necessary, the various constituent elements of expression: semes, which are identified by the dictionary; morphemes, identified either by form classes (if they are independent words) or by the stem-ending tables (if they are inflexions); syntagmas (expressed in formulae) identifiable by means of the inventory of the various possible structures of a given language.

It is clear that such a method cannot do more than facilitate the translation of word combinations of unambiguous structure. A phrase like "the King of England's Empire" will always remain enigmatic to the machine, since it contains no graphic or structural clue to determine whether it refers to the Empire of the King of England, or to the King of the Empire of England. In such cases a reviser must remain the only final resort, as in cases of lexical or morphological ambiguity not reducible to objective graphic criteria.

STRUCTURAL MEMORIES

Meanwhile, it is clear that the translation machine will have to add to the memories already described—lexical memory or dictionary, morphological memory or stem-ending tables,—a structural memory, permitting comparison of structures received at input with structures held in this structural memory. A comparison of this sort, while not of any great interest for simple sentences, will be essential for complex sentences and will make it possible to reduce to a minimum all cases of ambiguity inherent in the structure of the sentence itself. In human languages, it is syntax—the compulsory pattern for the combination of words into sentences—that is slowest to change. A complete inventory of the structures of a given language will require ingenuity in classification and finesse in establishing the order in which the machine is to conduct its explorations, rather than patience in drawing up the inventory itself. Such an inventory can be drawn up for each language and only when this has been done will fully automatic translation become possible. Compared with lexis, the problem is one which is relatively limited in scope.

For the same reasons, the problem of registering the rules of morphology and syntax in the memory of the machine does not, from the point of view of computer technique, present any problem more difficult to solve than those involved in programmes of management and scientific calculation already treated by machines. The data to be entered are not appreciably more numerous than the rules which must be stored for the execution of a sequential series of scientific calculations. A magnetic drum computer will probably be able to enter on its drum all the morphology and all the syntax of two languages—the only remaining questions in the field of technology being rapidity of access to the rules thus entered, and in the field of programming, the order of entry of these rules and of access to them.

CHAPTER VI

Lexical Problems of Automatic Translation

WHILE the problems of morphology and syntax are relatively restricted in number, the same cannot be said of questions raised by vocabulary. These are, indeed, very considerable in extent if not in complexity. Taking into account variations in the meaning of words, the rapid evolution of scientific and technical vocabulary, slang and local speech, the number of words per language may be so high as to challenge the skill of electronic memory constructors. In recent concise dictionaries the vocabulary of the English language comprises some 60,000 word entries: this number may run four times as high if each meaning of each polysemantic word is entered separately. So that a dictionary in which every form of every word would constitute a separate entry might well number over half a million words in a modern inflected language. We are thus faced immediately with the problem of lexical *content.* This is closely followed by questions of *classification* (should there be one dictionary only, or several, according to subject?), and of *order* of classification (alphabetical, logical, conceptual, or numerical according to the increasing or decreasing number of characters in a word, etc.?). Finally come the specific problems of translation—multiple meaning, idiom; and—sooner or later—the problem of style, or styles, of the choice of words for reasons peculiar to the author.

MEMORIES: TECHNICAL ALTERNATIVES

A solution to any one of these problems involves a choice, or a series of choices, inevitably limiting possibilities in other directions ; all the more so, in that lexical problems, even more than those of morphology, are closely bound up with the technological aspects of computer construction. Memory capacity, rapidity of access, these are important considerations in the choice of solutions. Even if we set aside for the moment, for practical reasons, the objections of those who maintain that the choice of the right word by the translator is a matter of taste, of personal judgment, and that the machine will never be able to exercise such judgment, we can still not affirm at the present time that an ideal solution has so far been found to the lexical problems of mechanical translation. But the empirical method of partial solutions has been applied with increasing success. It has enabled research to continue while technicians pursue the study of recording processes, thanks to which it will eventually be possible, where required, both to store a very great number of words and to have very rapid access to them.

One or more magnetic bands or a battery of magnetic discs can contain an entire dictionary: but access time to any given word is relatively long, and this means a slowing-up in the matching of input words with words stored in the dictionary. It has been calculated that the lexical memory of the machine should provide random access to any word in not more than 10 milliseconds; this would allow for the matching of dictionary words with all the words of a sentence of average length (20 words) in one-fifth of a second. Since tapes and discs can only give sequential access, it is the magnetic drum that at present appears the most suitable method of making vocabulary as well as rules of syntax and morphology instantly available for machine operations. The weakness of the drum is that its capacity is limited. Other types of memory with high capacity and rapid random access are, however, likely to be available shortly.

During the past two years Erwin Reifler and Lew Micklesen at Seattle have been concentrating on the linguistic work connected with the use of a large-scale Russian dictionary recorded on Gilbert King's photoscopic memory, which provides immense capacity and very rapid access, but the logical circuits necessary for translation have not yet been added to this memory. Similarly, at Harvard, Oettinger and others have been working on a Russian automatic dictionary on a Univac I computer. Both teams have made considerable progress in the lexical analysis of Russian and the logical treatment of vocabulary, and their methods, when applied to more modern computers, should lead to very rapid progress.

The work of Reifler and Oettinger clearly shows that there could be no excuse for awaiting the improvement of memories before beginning the basic linguistic work. All that is necessary is to ensure that lexical research is undertaken in the order best calculated to exploit the properties of existing machines while not losing sight of future potentialities. We may then hope that many of the purely linguistic aspects of the organization of lexical work will have been dealt with by the time the electronic technicians come forward with their optimum solutions.

Let us suppose that a large magnetic drum is used as support for a lexical memory. If the average length of a word is six letters (to which must be added grammatical indications and programme instructions appropriate to each word) we must allow for six characters × six bits per character + the indicators and instructions, that is in all some 250 bits per entry. Certain programmes may require as many as 1,000 bits per dictionary word. The magnetic drums of the type employed in the 704 computer now functioning in the I.B.M. Paris office store between them 294,912 bits—that is, according to the type of programme, between 300 and 1,200 words; the drum of the Gamma 60, containing 786,432 bits, has a capacity of 785 to 3,200 words. To execute a minimum programme a translating machine must therefore have a capacity at least as high as that of the I.B.M. 704 and may require twice this capacity unless it is arranged for the machine to have, for certain parts of the programme, rapid access to other types of memory which are called into play in certain cases only.

Several different types of lexical memory can be used. It is conceivable that one or more vocabularies comprising a very large number of words could be recorded on magnetic tapes, of unlimited capacity but sequential, and therefore slow, in access; that, for the special requirements of any given translation, one such vocabulary (or section of a very large vocabulary) could be transferred, for the duration of the operation only, on to either a drum or ferrites or any other form of memory providing rapid or ultra-rapid random access. In the course of one translation, it would then be possible, on receipt of a given signal, to call into play such and such a specialized vocabulary, registered on magnetic tape, to transfer it for a few minutes only on to a drum or ferrites, and to replace it some instants later by another similar vocabulary.

I TERMINOLOGY CENTRE

A. PERMANENT ELECTRONIC MEMORIES

Number of Stores: *1 complete set (1 to n+2) for each bilateral programme.*
Storage Medium: *varies according to needs, from (a) punched cards, (b) punched tape, (c) magnetic tapes or discs, to (d) photo-electric discs, etc.*

B. TECHNICAL EQUIPMENT

Punchers, sorters, magnetic recorders, tabulators, tape readers, etc.

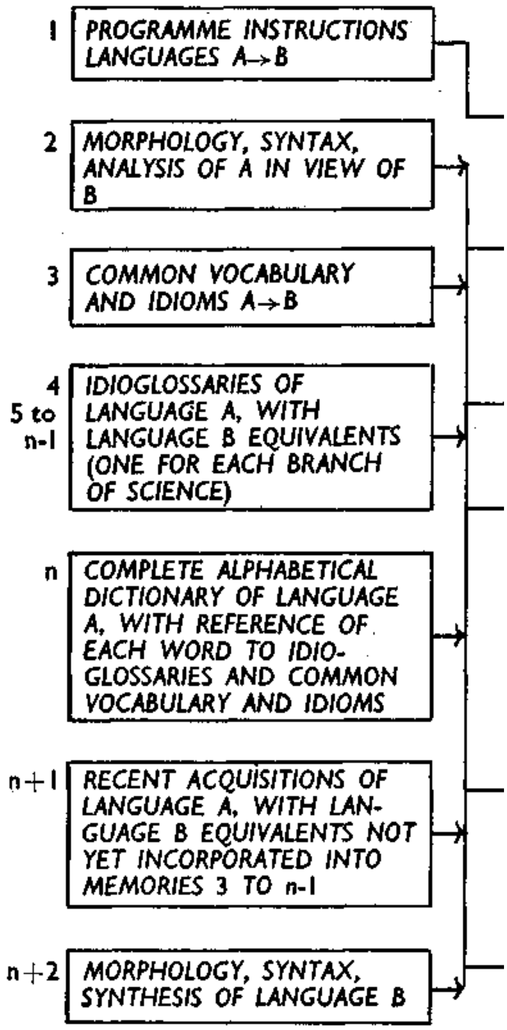
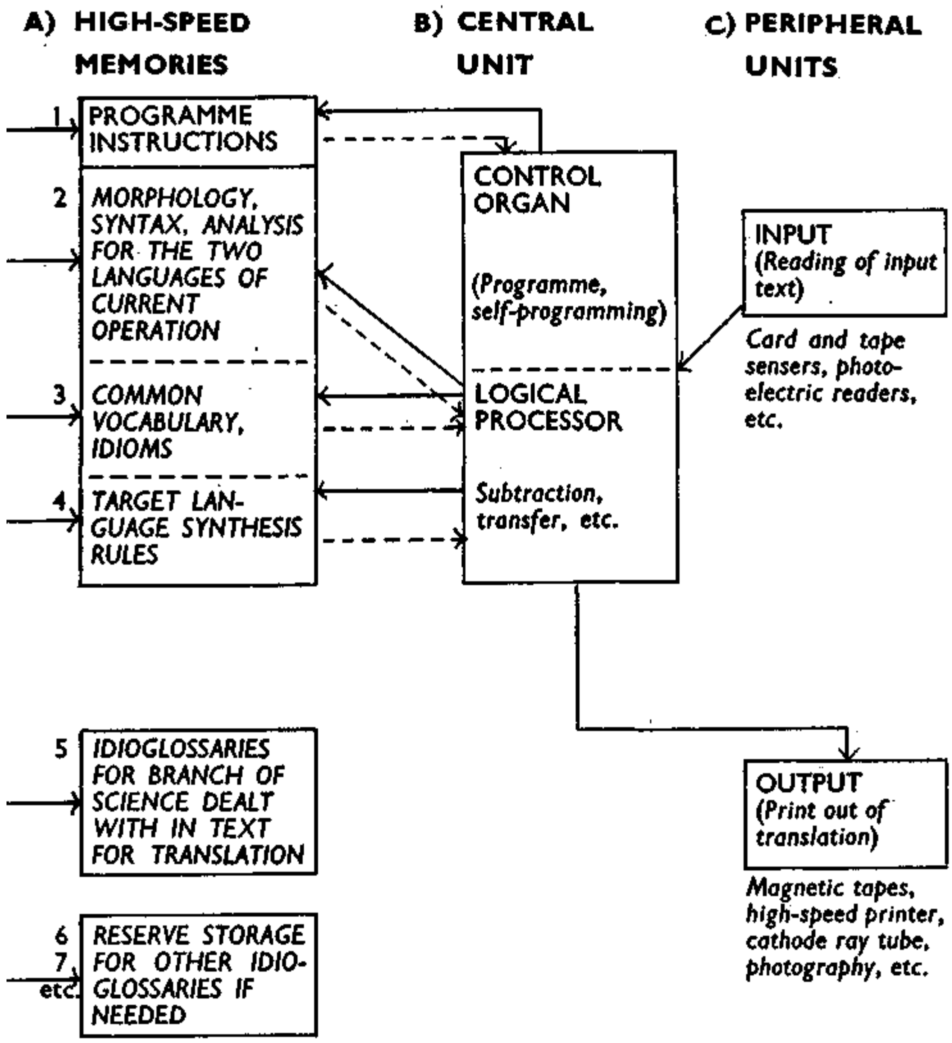


Fig. 4. Block-diagram of workshop including terminology centre and automatic translating machine.
The terminology centre and translating machine shown above have been designed in order to illustrate a hierarchy of lexical and other stores suitable for terminological reference work and for machine translation of languages. Machine operating cost has not been taken into consideration. Whereas this machine calls upon considerable resources of memory storage, the use of new programming procedures may well permit to operate on smaller-scale machines.

II TRANSLATING MACHINE



According to contents, elements 1 to 7 of high-speed memory may be recorded for operational purposes on one or more drums, one or more ferrite matrices.

The time of transfer being relatively negligible, a rational organization of vocabularies by subject is perfectly compatible with the simultaneous utilization of slow sequential memories and rapid random ones (see Figure 4).

CONSULTING THE ELECTRONIC DICTIONARY

The first question to be raised was how to classify words in an electronic dictionary in such a way as to ensure as rapid a look-up as possible. Words being represented by a binary numerical code, several alternative methods of classification have been tried: arrangement in order of decreasing frequency, alphabetically by sections, etc. Booth [7] has described the method recommended both by himself and by the Russians. "Suppose," he writes, "that the dictionary contains N entries arranged in ascending order of numerical magnitude in locations 1, 2 . . . , N and that N is some power of two. The incoming word is first subtracted from the entry in N/2. Then if the result is positive, the required entry is in the 'first half', i.e. between 1 and N/2. If negative, however, it is between N/2 and N. Now, assuming a negative result at the first stage, subtract the word from the middle entry of the last half, i.e. that in N/2+N/4. If the result is negative the equivalent must lie between entries N/2 and N/2+N/4 and, if positive, between 3N/4 and N. This comparison process is repeated until the correct location is isolated and it is seen that this requires $\log_2 N$ steps." With a dictionary of 10,000 words (10^4), 14 operations ($4\log_2 10$) are required; for 20,000 words (2×10^4) 15 operations, and for 1 million words (10^6) 20 operations. For a machine of which the combined access and subtraction time is 1 millisecond, the look-up time for one word in a million is about 20 milliseconds.

CODE COMPRESSION

A further technical refinement of great importance for the whole conception of the dictionary is code compression. In order to save memory space and thereby augment memory capacity, methods employed in cryptography and telegraphy have been adopted by mechanical translation mathematicians. One such method consists in adding together the code numbers of each six-letter group in the input word and treating the resulting total as representing the word.

LINGUISTIC PROBLEMS

Classification of words in ascending order of magnitude of their code and subsequent code compression are mathematical solutions to technological problems of recording linguistic data in the machine or of achieving greater speed of access. The real linguistic problems are no less urgent—for example the fundamental question of multiple meaning in relation to the dictionary. Should the dictionary contain as many entries for each word as that word has meanings? Would this drastic solution, which is perhaps compatible with a gigantic memory having very rapid access, be appropriate for a programme of sentence analysis based, as at present seems desirable, on the necessity for solving problems of multiple meaning by exploration of context? Or have we not rather arrived at the point where the complexity of programmes and the necessity of keeping such programmes flexible, argue in favour of restrictive dictionary size and concentrating on specialization by subjects or groups of subjects?

In all experiments to date certain precise limits have been imposed in order to achieve effective results without sacrificing the balance which it is desirable to maintain between the proportions of the machine and the relative size of the subject of the experiment.

A FRENCH-RUSSIAN DICTIONARY

Kulagina and Mel'čuk have constituted, according to this principle, an experimental electronic French-Russian dictionary for the translation of mathematical texts. Their method differs little basically from English and American electronic dictionaries. The texts of Paul Appel and Emile Borel on which the dictionary is based comprised 20,500 running words of which 2,300 were different. The 1,000 words occurring more than four times each were entered in the dictionary. Without any statistical survey, about 50 words that "were obviously needed" were added, together with another 50 French "grammatical tool" words. This gave a total dictionary of about 1,100 words. Each stem in the dictionary was accompanied by a dictionary entry containing:

(1) the Russian translation; (2) French data including *(a)* a part-of-speech notation, *(b)* an idiom notation, (c) the preposition code, *(d)* grammatical characteristics; (3) Russian data, including (a) a notation on selection of Russian stem, *(b)* grammatical characteristics; (4) a notation on the choice between two French stems. [18]

This method of noting the characteristics of each word is similar to that described by Panov and Bel'skaja for English vocabulary (see Chapter IV); in the case of French, the grammatical indications are, for example, for nouns: gender, formation of plural; for verbs: transitive or intransitive, conjugated with *être* or *avoir,* conjugation number, etc.

The preposition code corresponds to the peculiar problems presented by French prepositions which can be translated in many different ways, governing a number of different cases. This code refers the machine to special preposition translation tables. Preposition codes are given for nearly all verbs and many adjectives and nouns, the same preposition code number being given to all words governing the same preposition.

The notation on the "choice between two stems" consists of an indication that a choice of stems is involved and a notation giving the address of the alternative stem: for example *point,* noun and *point,* negative particle, will be accompanied by this indication so that the machine, having completed the look-up and the processing of idioms, can, by using the rules for distinguishing homographs, "decide" which of the two stem entries applies in that particular case.

IDIOMS AND HOMOGRAPHS

The indication "idiom" means that a word may, in association with others, form a group, the meaning of which is not dependent on the analysis of each word in the group, so that a literal translation would either be meaningless or would convey a wrong meaning. This indication refers the machine to a special dictionary, where all idioms containing the word bearing that indication are listed. They are divided into integral (e.g. *de plus en plus, à présent)* and non-integral idioms (e.g. *aussi . . . que, à . . . près)* arranged in the alphabetical order of the meaningful word of each idiom. Under the same meaningful key word, which may occur in several idioms, they are arranged in decreasing order of the number of words in each idiom. A special indication in the stem dictionary gives the number of idioms listed under each key word, so that the search programme may come to an end when the list is exhausted.

Thus when the machine finds the word *plus,* the stem dictionary refers it to the idiom dictionary. If the machine identifies a group of input words (for instance *de plus en plus)* with one of the idioms listed, it finds the translation; if not, it returns to the word *plus* and translates it in accordance with the instructions of the stem dictionary.

A distinction is thus established between the analytical constituents of language—which the speaker is still free to combine as he wishes—and fossil or vestigial constituents, which, while they are not single words, are nevertheless units of meaning which can no longer be analysed. This distinction shows up the true nature of idiom in our modern languages: as fossilized survivals of expressions which were originally analytical.

Idioms do not, in scientific language at least, present a problem of any great magnitude. It is necessary only to catalogue them and record them in a special memory. The problem will doubtless be very different when we come to everyday language and to that of plays and novels. Since the use of idiom introduces an extra-linguistic element into language—the evocation of a situation which has a special meaning for a given social group, a systematic study will have to be made of idioms, clichés and all metaphorical use of words or groups of words—and Flaubert's original idea of a *Dictionnaire des Idées Reçues* will perhaps enjoy new popularity and expansion! A study of this type would make it possible to decide which idioms form part of everyday language, and must therefore figure in the idiom dictionary; which can be translated literally into certain languages, and which must perforce be left in the original language as being totally untranslatable since they refer to social situations of uniquely local and limited significance.

The problem of homographs is relatively restricted. We have seen in the previous chapter how grammatical analysis of context enables the machine to choose the right translation for most homographs. In the rare cases where such analysis is insufficient, they will probably have to be classified with genuinely polysemantic words.

GENUINE POLYSEMY

There are many words, apart from those with idiomatic usage and those whose meaning varies with their grammatical function, which are in fact truly polysemantic. Is the English *plant* a French *plante* or *usine?* Is the French *temps* to be translated *time* or *weather?* Should *champignon* be rendered by *fungus,* by *mushroom* or by *toadstool?* Grammatical analysis is of no assistance, nor at first sight is the idiom dictionary.

What does the translator do when faced with such a problem? If he understands the subject perfectly he chooses the translation which appears to him to correspond to the overall sense of the context. In a sentence dealing with poisoning, he will translate *champignons* by *toadstools*—although he will be understood if he says *fungi.* But scientific and technical translations are full of traps for the human translator not fully conversant with the subject of his text. Only constant and close collaboration between translator and specialist will ensure that the right translation is always given—above all in texts on modern technical subjects where the vocabulary is in constant evolution.

The translation machine cannot hope to do better than the human translator in this respect; if the text fails to provide the machine with recognizable, objective criteria signalizing meaning, then the translation will, for the present, have to list all the meanings of indistinguishably polysemantic words, and a specialist will have to choose the right meaning from this list before the final version is made. It is, however, obviously desirable that the machine should be able to solve the majority of polysemantic problems. It should be able to choose the right meaning. In cases of grammatical multiple meaning and of homographs, we have seen that the micro-context—the study of the immediately surrounding words—has made it possible to choose automatically between several meanings.

How can the context help to determine the correct English rendering for a polysemantic word like *champignon* or to decide on the correct English equivalent for *temps?* It is possible to imagine a general dictionary for a given language A containing a translation for every single word in language B. A word would be defined as a meaningful group of signs, alphabetic and nonalphabetic (spaces, hyphens, etc.) which have one meaning only. A word M with four distinct meanings in language A and requiring four different words for translation into language B, would thus figure four times in the dictionary, as M1, M2, M3, M4, with appropriate indications making it possible to identify with one of these four meanings, according to the context, its correct translation. But for this it is essential that the sentences to be translated should contain objective criteria enabling the machine to choose between M1, M2, M3, and M4, and only extensive analyses of context will make it possible to see to what extent they so do.

Polysemantic words which have exact polysemantic equivalents in another language are of no great importance: the overall sense of the context will provide the reader with the means of choosing between the four meanings of a given English word provided these four meanings coincide exactly with the four meanings of a given French word. Problems arise where the multiple meanings do not coincide between two given languages, that is to say where there are differences between the connotations of a word in language A and those of the word which normally translates it in language B.

MICROGLOSSARIES

Of all the solutions so far suggested, the most practical seems to be the idea of idioglossaries or microglossaries. It had originally been thought that by listing in the output translation all possible meanings of a word in language A, the reader could select the correct one according to context. Research since 1949 has led to the provisional conclusion that in scientific texts non-grammatical polysemantic nouns and verbs do not present any great difficulty within the limits of the restricted vocabulary of any given science or technical subject. Thus special restricted dictionaries—microglossaries—should be constituted, having the double advantage of reducing the size of the dictionary necessary for a given translation to dimensions compatible with the operational memory of present-day computers, and also of limiting the number of cases of non-grammatical polysemantic words.

The Academy of Sciences of the U.S.S.R. considers a dictionary of 6,000 words quite sufficient for translating any mathematical text. They think it reasonable to expect that other fields will not require much larger vocabularies. This estimate is borne out by the statistics showing that 95% of English texts can be understood by a reader knowing 6,000 words. The specialized mathematical dictionary established at the Academy for the translation of Milne's *The Numerical Solution of Differential Equations* was divided into three independent sections:

(1) Technical words, i.e. mathematical terminology—approximately 400 words.
(2) Non-technical mono-semantic words, amounting to 1,800 words.
(3) Polysemantic words, amounting to 300.

The technical words of a subject being thus recorded in a special memory, it becomes relatively easy to find their exact translation *for this subject,* it being assumed that multiple meaning is rare within the limits of one scientific subject. The one remaining problem is that of multiple meaning of words which have one meaning in mathematics, for instance, and another in physics, in a sentence dealing with both mathematics and physics. Here the machine is at a disadvantage compared with the specialist translator, but not greatly so compared with the non-specialist.

Andreev suggests determining the particular meaning of a polysemantic word by a system of "semantic keys", of the type employed by lexicographers for identifying particular acceptances of words. In an article on agriculture, for example, the word *luk* in Russian has every likelihood of meaning *onion* and not *bow;* in an article on astronomy, *vozmuščenie* will almost certainly mean *perturbation, a change in the orbit of one celestial body under the influence of another,* and not the mental state *indignation.* While secondary meanings cannot be absolutely excluded in such cases, their probable incidence according to Andreev [1] is close to zero. "Hence the percentage of errors resulting from disregarding the secondary meaning will in general not be greater than the usual percentage of typographical errors." When receiving the text for translation the machine will be provided with a semantic key permitting it to select immediately from a general dictionary the particular meaning of a polysemantic term corresponding to the subject of the text. This, of course, would be particularly suitable in a large-size dictionary such as that on which Reifler has recently been working in Seattle.

Andreev further recommends sub-dividing the dictionary into separate fields: mathematics, chemistry, zoology, music, etc., each with its own semantic key or numerical code establishing a relation between a word and the subject of the appropriate section. The translation will proceed by successive look-up operations in the different sections of the dictionary, beginning with the main subject, i.e. the mathematical section if the text relates to mathematics, and so forth. The general dictionary will be consulted only after the special sections. It will itself be sub-divided according to indications provided by a statistical study of vocabulary, into the following sections: (1) commonly used words, (2) words of average frequency, and (3) rarely used words. Dictionary search will proceed in the numerical order of these three categories, and only words which are not found in the first category will be looked for in the second and so on. An appreciable amount of time will thus be saved in dictionary look-up.

STATISTICS OF WORD MEANINGS

All these various solutions to the problem of genuine polysemy involve a pre-selection of the meaning of words by means of the macrocontext, while use of the microcontext makes it possible to solve problems of grammatical polysemy and to identify idioms. Both approaches are founded on a probabilist attitude towards the problem of meaning: it is therefore essential to make thorough statistical studies, based on numerous and varied texts, of the exact meaning of words in language A and their equivalents in language B. Such a study should be focused not only on the frequency of words as represented by alphabetical signs, as are those of Zipf and Estoup, but also on the frequency of the various meanings of polysemantic words. Macrolinguistic analysis will be brought to bear on sign/meaning combinations and not on signs alone.

Linguists who become automatic translation programmers will have to be trained in probabilist methods. If a word has a certain meaning in 95% of cases and alternative meanings in 2% and 3% of cases, it may be necessary to risk translating it by the first meaning and to give the other two only in parentheses, or to ignore them altogether. The bilingual dictionary of automatic translation will be based on this principle. Such calculated acceptance of risk is also necessary in organizing human translation, which is never altogether devoid of erroneous shades of meaning: as in typography, it is the low percentage of error which determines the quality of the work—total absence of such errors is very rare. It is probable that by making systematic inventories of vocabularies and synonyms it will be possible, for scientific texts, to isolate a relatively restricted number of cases where only specialists in the subject will be capable of determining the correct translation in a given context.

Reifler, in one of his studies, rightly emphasizes the importance of comparative semantic studies for the eventual reduction of the role of the reviser of automatic translations. One aim of such studies might be to determine which word in language B translates most completely all the meanings of a word in language A. This translation would then be adopted as the most satisfactory from the point of view of the reader's comprehension of the meaning of the original text. For instance the English *fungus* could always be used to translate *champignon,* as being the only word communicating the total connotation of the French word. Thus automatic translation, for the sake of communicating meaning, would seek a simplification and concentration of vocabulary, similar to that observed during all the great classical periods of history, when meanings with wide connotative values in the given community are preferred to individual and local semantic fantasies.

How far should such studies be carried? We can do no more here than to suggest a few of the avenues open to us. The question has been asked, how can the machine distinguish between the meanings of the French *temps, time* and *weather?* If we turn to Littré and see how many times *temps* suggests the passage of time and how many times it refers to sun and rain, we find that the rare examples quoted by Littré where *temps* means *weather* can all be considered colloquial or idiomatic, the special meaning being identifiable with the help of objective criteria present in the micro-context. The study of idiom and that of the particular meanings of certain words leads straight into the field of comparative and statistical semantics, and it is here that the key to many poly-semantic problems seems likely to be found.

We thus arrive at the need for a taxonomy of the meanings of words, and for logical definition and classification of concepts and the ideas which express them—to the synchronic study of changes in meaning by metonymy and synecdoch.

THE *THESAURUS* METHOD

That is broadly the direction taken by the British team working in the Cambridge Language Research Unit under the direction of the logician Margaret Masterman. They have had the idea of solving problems of meaning equivalence between languages by a device based on the conception of Roget's *Thesaurus of English Words and Phrases,* in which words are classed according to the ideas they express. Ideas are classified logically, generally in dichotomic form, under numbered headings. Thus Roget had established a concordance between a logical classification of concepts and a numerical system which can be processed through a computer. The same English word with several meanings will appear under several headings, accompanied by other words expressing the same or closely related concepts, and thus narrowing down the possible meaning.

In order to reach the exact language B equivalent of a word in language A, the Cambridge research unit have thought of improving on the *Thesaurus,* and of searching by machine for the word in language B which is common to all the *Thesaurus* headings under which the words in the immediate context of this word can be found. The method is attractively ingenious but proves somewhat disappointingly clumsy in application. Numerous systematic trials alone can show whether it can be of real service. It seems, however, open to a fundamental criticism: is it necessary or useful to look for the correct word *in the output language,* by a method fraught with hazards? Is it not simpler to look deliberately in the electronic dictionary for the translation of a given word, having used the context in the original language to define the meaning of that word? However seductive and original some of the ideas of the Cambridge unit, it seems paradoxical and contrary to the necessary respect for the intentions of the writer, to place the emphasis on the output language in lexical search for the right word. The philosophy of thought and of its means of expression which is behind the work of this group derives from the disciplines of logic rather than of linguistics and psychology, and is therefore somewhat divorced from the empirical approach which the problem of translation requires.

Nevertheless, this Cambridge research leads in the direction of a new kind of bilingual or multilingual dictionaries, in which words would be classified according to the ideas they suggest, with a numbering system referring to their logical position within a taxonomy of concepts. This would be a refined variant of Andreev's suggestion, based not on notions of simple frequency but on the relationship between words and a logical classification of ideas. The *Thesaurus* method, applied not to the operations of the machine at the output stage—where it would run the risk of overlooking certain fundamental requirements of translation—but to the classification of words in the dictionary, would probably facilitate the search for exact meaning in certain cases of polysemy by bringing the immediate context to bear more completely on the polysemantic word.

SCIENTIFIC AND TECHNICAL DICTIONARIES

Apart from the thousand or so common words indispensable to any translation, the vocabulary of science and technology will be the first to be subjected to terminological classification for mechanical translation. It constitutes a high priority field in view of the urgent needs of science, and a particularly propitious one in that words generally have only one meaning within the limits of the glossary of a single scientific subject.

The lexicography of mechanical translation is now aiming at the construction of specialized glossaries magnetically or otherwise recorded. The work of Oettinger at Harvard, Panov and Bel'skaja in Moscow, Reifler in Seattle, shows the way. But, as Oettinger's latest report proves [29] the construction of a dictionary is inseparable from work on morphology and syntax and from the many aspects of linguistic research involved in and facilitated by computer analysis.

Like the Harvard team, the research workers at the Rand Corporation of Santa Monica in California have undertaken under Kenneth Harper and David Hays a systematic inventory of words and rules of translation—in spite of all attempts at analysis, the two remain indissoluble—and have defined the guiding principles of their work. The inventory begins with 20,000 to 50,000 running words taken from one or more scientific texts in one language and dealing with one subject. These texts are systematically analysed by means of punched cards and electronic machines. Morphological and syntactical rules with a bearing on translation are inventoried; the vocabulary is classified into monosemantic and polysemantic grammatical words, common words, and words peculiar to the subject of the text. Note is taken of terms which, having a different meaning in other contexts, are to be the subject of separate study. A translation is then made and treated by the same procedure of word and context analysis. Thus we have a first vocabulary of words in common use in each of the two languages under examination, plus a bilingual dictionary of the scientific subject chosen. A second corpus of text is translated mechanically by means of the vocabulary and rules drawn up from the first translation; this new translation is revised and improved where necessary, the decks of punched cards are completed in accordance with those improvements wherever necessary and the whole process is then repeated. The same operation, carried out as often as necessary on other texts as closely as possible related in subject to the first text, will make it possible to constitute step-by-step a complete bilingual vocabulary of this scientific subject, the common vocabulary being both enriched and defined in the process. For this purpose a corpus of some 250,000 to 500,000 words on the same subject would be subjected to analysis by this cyclic process. Acceptable translations are a by-product of this method, clearly illustrating its value as a means of accumulating objective knowledge of language.

By moving on to another subject—proceeding from mathematics to a branch of physics or astronomy, for example—the bilingual scientific vocabulary is gradually extended, and a series of microglossaries can be created in which the enrichment of the vocabulary goes always side-by-side with the statistical analysis of results: word frequency, incidence of exact language-to-language equivalences, nature and frequency of polysemantic words, etc.

A NATIONAL TERMINOLOGICAL CENTRE AND TRANSLATION LABORATORY

This method, inseparably bound up with ultra-rapid electronic recording processes, contains in embryo the solution to the basic problems of scientific and technical translation, whether by man or by machine. In the first place it will permit the constitution of technical dictionaries which can be constantly kept up to date and available for rapid and reliable consultation. We are, however, dealing with collective means of production, the complexity and costliness of which will make it necessary to operate on a nation-wide or even international scale.

When once an inventory has been made of all the Russian words, for instance, used for one technical or scientific subject and their French equivalents, and it has been magnetically recorded in binary code, a national electronic centre for terminology could continue to receive from specialized research bureaux, in the form of typewritten or better still punched cards, both requests for technological information and new acquisitions in technological terminology. The electronic dictionary for any given scientific subject would thus be kept up to date at regular intervals by the automatic insertion of new words or meanings on magnetic bands, which would, in due course, be used to “charge” the magnetic drums or ferrites of the rapid access memory of translation machines. These magnetic tapes would serve a dual purpose: not only would they constitute the permanent technological memory of translating machines, but they could also be "read" in reply to questions received from research centres, and this “reading” would enable the reply to be punched or typed on the incoming cards, which can then be returned to the bureaux from which they came. A central automatic translation workshop would thus necessarily be at the same time a national or regional focal point for terminological information and would replace or complete lexicographical cards of the conventional type (see Figure 4).

FROM METALANGUAGE TO THE UNTRANSLATABLE

In what order should this work of drawing up a terminological inventory, of cataloguing the sum of human knowledge, be undertaken? The facts dictate a certain order. The specialized vocabularies should be explored in the order defined more than a hundred years ago by Auguste Comte, that of the decreasing exactitude of the sciences. The effort required will increase with the decreasing degree of precision of each branch of knowledge. It is no accident that has led the Soviets to begin with mathematics. Astronomy also brings to automatic translation clear and distinct concepts, an international terminology free of individual fantasies, analogies and figures of speech, those traps set by non-Cartesian thought on the path of all translation which seeks to be exact and faithful. Next will come the natural sciences, in their descriptive and mathematical aspects, the concepts of which are often expressed in formulae or in exact definition of objects; their terminology will be all the simpler to catalogue in bilingual form in that their vocabulary describes substances, recognized facts and relationships which can be perfectly expressed without any admixture of metaphoric language.

With the human sciences, social and psychological, the problems become more complicated and the lexical work in particular becomes extremely arduous, since it must bestride two languages and yet take into account the semantic variations introduced by individuals in the use of words and in the creation of vocabulary, inevitable when new ideas are to be expressed. These sciences are already far removed from formulae and metalanguage, and are a fertile field for image and figures of speech.

Petroleum research engineers, for whom a “carrot” is the contents of the boring tool taking soundings from the subsoil, atomic scientists for whom the atomic pile which generates fissionable isotopes is a “breeder”, face the translator with problems as hard as those set by the statesman making a speech at the United Nations who interlards his words with proverbs and images drawn from his national folklore, without equivalent in any foreign tongue. These men are thinking in terms of imagery, not in exact definitions. Brilliant improvisations translating these images will be much admired and long remembered: only the untranslatable really requires translation. You do not need to translate H_2O and $ax^2 + bx + c$. The more closely language reflects the reality it expresses, the less it is necessary to translate from one language into another. It is metaphorical neologisms, not the clear and distinct ideas of the sciences, which cause disputes in terminology centres and offices for the standardization of technical terms. They too necessitate the most laborious research on the part of the translators of international organizations. The art of the translator begins at the point where thought diverges from the descriptive and analytical methods of the sciences, into the twilight zone where judgment and feeling play as great a part as knowledge. In this area successes are sometimes easy to score, but traps are many, disputes often violent and personal preferences vehemently expressed.

This is also the zone of genuine polysemy, which taxes equally machine and translator, since the lack of objective criteria is bound to leave open several alternative choices. No such criteria are available to guide the translator faced with the metaphorical use of words (like *carrot* or *breeder)* such as are constantly found in technical texts, and in the writings of sociologists, economists, psychologists, psycho-analysts and even linguists, all of whom tend to confuse language the tool of their analysis, with language the object of their study, because the subject of their work has no material being other than in words.

In such cases the translating machine, in spite of its special requirements, can be of some service. The author of obscure texts, in which the slippery shifts of semantic meaning are never signalized, is always inclined to blame the translator for not following every fluctuation in the meaning of the words he uses. Will he be able to blame the machine if he has used the same word in two different senses? The machine will be able to translate writings on the human sciences only after thorough preparatory work on their terminology. In this respect it will not differ from its human predecessor. Such preparation will be facilitated if authors will consent to define their terms and to add a glossary to their books and articles.

STYLES AND VOCABULARIES

Nor do the prospects opened up by lexical research for automatic translation machines end here. Certainly it will be relatively easier for the machine to translate scientific and technical works—and these are at once the most urgently needed and the most economically rewarding translations. But the very logic of the work of programming for the translating machine may lead on to bolder enterprises.

Is there in fact any definite and firm line of demarcation between the translation of scientific and technical prose and that of literary prose? Where does scientific vocabulary end and that of literature begin? If, beginning with scientific prose, it is necessary in order to solve the problems of automatic translation to break down the vocabulary into different compartments so that the memories of the machine can more easily digest it, this same method is capable of extension to all vocabularies until all possible groupings of words by subjects, or indeed from any other conceivable angle, have been exhausted. The lexicography of the translation machine thus leads finally to a general logical classification of knowledge, that is to say of the words expressing knowledge, the overriding law of such classification being that of commodity of access to the data necessary for the translation of any given text. The principal, if not the sole superiority of strictly scientific texts over those of general literature probably consists in the stricter limitation of the theoretical scope of their vocabulary, in a more exact equivalence between the objects described and the words used to describe them, that is in the higher quality of their information content due to the more exact definition of the meaning of each term.

But these are differences not of kind, but of degree. Beyond the first subdivision of vocabulary by scientific disciplines, how far will it be practicable to pursue the ramifications of a general classification of all words, by subjects, areas of geography, periods of history, social environments, etc.?

Only a thorough study of the best techniques of storage and rapid access can reveal what is practicable. But theoretically at least, as long as we keep to legitimate subjects of study and translation, the possibilities are limitless. Geography, history, anthropology, ethnology will each contribute their specialized vocabularies and the terms used to describe the objects of these sciences will be infinitely numerous and susceptible of varied cross-referencing. In the same way that any good translator makes his own dictionary for each new author or subject, a great deal of the time of automatic translation programmers will be devoted, once relatively simple texts have provided the groundwork, to the selection of those different compartments of the electronic dictionary which must be called into play for any given work.

Shall we compile a special sixteenth- or seventeenth-century dictionary for the best historical novels? An Anglo-French dictionary of the works of Victor Hugo? A German-Japanese one for

Goethe? Will there not be room for an English-French Shakespearian glossary, and so on? In translating the great authors will it not be useful and perhaps even practicable to identify the exact meaning of the terms employed by them at the various epochs of their life? We are not here in the realm of science fiction; these are practical possibilities for the day, which is not far off, when critics, men of letters and literary translators come to make use of machines and magnetic recording systems. The present writer is gratified to find that on this as on other related subjects his personal views have led him to conclusions very similar to those of Professor Panov and Miss Bel'skaja. [25, 29]

THE SEMANTIC ATLAS

Lexical research for the translating machine opens up the way towards a vast collective study of vocabulary, towards the enumeration and classification on a national scale of all the words of a language, arranged, in order of frequency and date of occurrence, in specialized compartments of the electronic dictionary which can be called into play as required by each translation programme, or consulted like an ordinary dictionary on particular points of detail. This lexical work would make it possible for research workers to accomplish for the vocabulary of any language what Gilliéron has done for French phonetics—to draw up an atlas of meanings. A history of the changes in the meanings of each word should also be possible. If each nation, each linguistic group, participates in this study according to methods defined by mutual agreement, language-to-language equivalences of meaning can be patiently surveyed and recorded on magnetic tape, until they constitute a collection of bilingual electronic dictionaries which will enable automation to be applied progressively to an increasing number of languages and subjects and to the literatures of the present, past and future.

CHAPTER VII

Future Prospects

LIMITATIONS OF THE MACHINE

WITHIN limits, automatic translation is already possible; all that is required is that sufficient time and talent should be devoted to the preparation of bilateral programmes. Despite differences in theoretical approach between various schools of thought, success will be achieved provided that, according to the rules of scientific empiricism, all theories are turned to account.

It is now certain that the machine can transpose into a second language, correctly—that is to say respecting the rules of grammar and of syntax—a sequence of sentences written in an original language. Naturally it will not at first be able to avoid displeasing repetitions of the same word; it will not clarify ambiguities in the original text; it will not always avoid facing the reader with a choice between several alternative translations of a single word. It will have no particular style, or, if it has, that style will be a somewhat simplified style, that is to say it will transpose faithfully sequences of words or groups of words without seeking those short cuts, paraphrases and euphonies which a good translator who “rethinks” the original will always find. The degree of semantic sophistication of the machine will depend on that of its electronic dictionary; it will correspond to the degree of complexity permissible in the lexical programmes of the machine, which means, in the last analysis, on the number of numerical indices by means of which it is possible to determine useful choice between several alternative meanings of a word without unduly burdening and slowing down the programme.

ROLE OF THE MACHINE

Within these limits, what then is to be the role of the machine? In the field of scientific and technical translation the automatic translator will clearly be increasingly useful as and when the vocabularies of the various branches of science are inventoried and recorded in bilingual form, according to accepted norms of bilateral automatic translation. Already Oettinger's Harvard automatic Russian dictionary is proving its usefulness. The number of semantic gaps will decrease as programmes become more sophisticated and as bilingual vocabularies are completed. Such gaps will be of two kinds: words not yet appearing in the dictionary—which will be reproduced in the translation in their original form—and rare idioms not yet registered in the "idiom dictionary". Such idioms will be translated word-for-word, a solution which, according to Bar Hillel, will prove acceptable in a surprising number of cases, when what is required is communication of meaning and not a sophisticated rewrite bringing into play all the arts of persuasion.

The circumstances in which the meaning of a scientific statement is communicated to a specialist are, in fact, somewhat peculiar. Thanks to their intimate knowledge of the same subject, reader and author already enjoy, in most cases, a certain communion of thought. The art of persuasion, rhetoric, plays only a minor role in this meeting of minds. While the rough translations provided by Booth and Richens in 1952 were already of some slight value to scientists, those of Dostert and I.B.M., of Panov and Bel'skaja, of Oettinger and the Harvard group, are largely adequate for current scientific needs.

Moreover, it is possible, and probably essential, to foresee several grades of presentation for texts translated by the machine. Although it is true that translation will not be truly automatic unless we can reduce to a minimum the role of the reviser improving the translated text, nevertheless, we cannot accept without reserves the disappearance of the post-editor. The machine can dispense with his services as regards the syntax and grammar of the target language. Its raw output can be communicated to the scientist, to a restricted circle of interested specialists. But if this same text is to appear in a learned journal, or to be shown to a meeting of company directors, it will probably have to be touched up in order to reply in advance to the objections which the purists will not fail to raise. If it is to be more widely diffused, then it is certain that it will have to be carefully edited. And this is, in fact, exactly the procedure adopted for conventional translation.

OPERATING COST

Clearly the great advantage of the machine lies in its speed of operation: reliability will be a second advantage, once the vocabulary has been established. Well-trained and experienced translators, whose translations nevertheless require revision and editing before presentation to a relatively exacting public, normally translate 300 words an hour, counting the time required for research and careful preparation for the work. Even these good translators are liable to distort shades of meaning when not fully conversant with every detail of the subject. It is estimated that, provided they have been adequately programmed, existing machines could translate at the rate of 20,000 words an hour. Such a speed of output will require long and painstaking preparation and will demand a considerable investment in human effort and intelligence. Once achieved, this output will increase with the greater potentialities of machines now in preparation and about to be put into operation. Not only will this investment be amortized over a considerable number of years, since it is a permanent one—to all intents and purposes indestructible provided elementary precautions are taken—but the effort of preparation will preclude, or at least reduce, technical errors of meaning.

Moreover, good translators, of whom there are not enough at the present time to satisfy scientific requirements, will be employed either for the preparation of the lexical programmes of the machine or as revisers. Their productivity will be increased; once again mechanization of the purely repetitive elements of a complex activity will concentrate attention on other elements of that same activity requiring intelligence and invention. The net gain for science will be to render rapidly accessible works which today are available without undue delay only to those specialists who are also linguists: the division of labour in scientific research will be improved—hence new possibilities of creative thinking and cross-fertilization of minds.

Will the translating machine ever be a paying proposition? This will obviously depend on the speed of execution of its programmes and on the use made of these machines once the programmes are established. The most recent estimates forecast machine translations at a price definitely lower than the present cost of scientific translations. An American estimate quotes a maximum cost price of the order of $.005 (half a cent) a word; a more recent English calculation puts it at a maximum of 2*s*. a thousand words [16] the cost of automatic translation of scientific texts. Outside translations of such texts cost at present up to £3 10*s*. 0*d*. per thousand words, counting the incidental overhead expenses of commissioning outside translation. The English author quoted puts at £4 2*s*. 0*d*. a thousand words the price of outside translation, including administrative overheads. So that a cost of $5 a thousand words for mechanical translation, unrevised but technically perfectly correct, would be well worth while, above all taken in conjunction with the high speed of the machine which would greatly reduce the time lag. A cost of 2*s*. a thousand words would mean a sensational saving, even if the price of the preparation of the programmes and a normal revision fee had to be added.

Apart from all question of purely commercial values, automatic translation of languages brings appreciable advantages to the linguistic group or national community. Once the initial effort has been made and programmes established, it should free intellectual ability for more productive work than that of run-of-the-mill translation. Just as accounting machines perform mechanical work formerly done by men, the translating machine will assume the worst drudgery of the sometimes somewhat sterile business of translation. The translator is often a man capable of invention, of literary creation, of understanding subjects the complexity of which requires a high level of general culture. Think for a moment of the time such a man must devote to transposing from one language to another the personal pronouns, definite and indefinite articles, prepositions, conjunctions, everyday words and auxiliary verbs. When we read sentences, which the translator must translate from beginning to end, let us stop to consider the idea of *redundancy* in language with which the mathematical theory of information has made us familiar. Of all the words in a given sentence, how many are essential to the transmission of the author's message and how many are simply conventional signals forming part of the linguistic mould of thought, but not of the actual thought expressed by the author? If the whole work of translating this mould can be mechanized, and if, in addition, the automation of translation processes can be brought to bear on all or almost all those words of which the meaning really matters, may we not then expect to see a significant release of energy and talent? The tool now being forged will soon become an indispensable part of the intellectual equipment of every nation and its use should speed up the rhythm of the acquisition of knowledge and lead us to a wider and more equitable distribution of enlightenment.

LITERARY PROSE

Is it too soon to envisage the extension of its use to tasks normally considered as literary—the translation of general information, of books of travel and geography, of novels, of philosophical and critical works? Which of us has not in the past read translations of foreign literature, under the auspices and even under the signature of well-known authors, in which wrong shades of meaning and misconstructions have abounded because the translator—or his hack—had translated the words without understanding their meaning, or failed to recognize idioms, relying haphazard on his dictionary, or worse still, on his own intuition? Though frequently of blatantly poor quality, literary translation, when it plays its proper role, serves to build a bridge between different cultures. What counts is the imaginative effort to transpose not words, but representations. This is a supra-linguistic effort, delicate and complex in that the details of the representation of the reader rarely coincide with those of the author as between one culture and another. The role of the translator is to establish between the two a zone of intercommunication bounded by the evocative value of words. Here we enter the domain of supra-semantic evocation, of subconscious association, of harmonics and the magic power of words. Are we bold enough to trespass with our machine into this sacred realm of the individual and the imponderable?

Let us imagine that it is desired to translate into French a contemporary novel written in Hindi, the action of which takes place in a village in the Punjab. If the translator has in fact participated in the life of such a village and if, therefore, all the words of the original have for him their full local evocative power, he will at once find himself faced with the problem of translating the everyday vocabulary, not to mention for the moment the philosophical and religious undertones and the rhythmic and euphonic values of the original. Is he to speak of the *"maire"* and the *"garde champêtre"*? Should he call the houses, the familiar objects, cows and horses, officials and trees by their Hindi names transcribed into French, or should he seek equivalents in the everyday vocabulary of France? Neither of these two solutions is without its drawbacks: to take the reader completely out of his own element renders understanding impossible, whereas to gallicize everything is completely to destroy all local colour and feeling.

Similar problems must have faced the first translators of Russian novels, and it will be remembered how much the solution of partial translation left to be desired. In reality, we are dealing here not with translation pure and simple, but with wide cultural exchange and with the interpretation of one culture to another. We have to create for the French reader an atmosphere which he is only partially prepared to perceive, to take him out of his own element without permitting him to get lost: we must enable him to follow the thread of events and feel at least a part of the inner meaning of each scene. The brunt of this delicate mission will fall upon vocabulary. Only by weaving and reweaving the threads of his translation will the translator succeed in finding the right proportion of Hindi words to convey local colour and of French words to facilitate the adaptation of his reader to the change of scene.

How far will it be possible to "pre-fabricate", so to speak, this vocabulary when preparing a programme of automatic translation, by establishing in advance a mixed vocabulary peculiar to such a translation? There is nothing inconceivable about such an operation, which might even prove to be a paying proposition for a work of considerable proportions or for a series of works of lesser size. The machine would in no way take the place of man, but would perform for him certain ultra-rapid tasks in accordance with directives determined by the translator, who would then improve the detail of the machine's rough draft, as does any good reviser of human translations. The machine would simply have "devilled" for the man, but would no doubt prove to be more docile than a human hack and would lay fewer traps for the reviser in the form of wrong shades of meaning and plausible mistranslations. All is here a question of proportion, of common sense and of comparative costs.

COLLECTIVE METHODS

We shall perhaps see the modern equivalent of certain successful translation teams of old reconstituted round the machine. But the work of translation will be distributed somewhat differently: the machine will replace the hack translator, while part of the team concentrates its energies on thorough preparation of vocabulary and others are detailed to work on revision and stylistic improvement of the translations coming from the robot at great speed.

Here again the machine leads to collective methods of work, by concentrating and accelerating certain means of production which in the modern world are rarely suited any longer to individual work and are therefore best used collectively. Electronic memories, by their reliability and their speed of reference, make it possible to devise methods of translation which will associate the best specialists of languages and of the sciences, each collaborating in the common task and contributing to it his individual knowledge. The collective methods of research laboratories are those needed for this literary and artistic work. Thanks to electronic memories this collaboration will be effective not only now, but will carry over from one generation to another, as does the work of the great lexicographers: each programme, each dictionary, once established, can serve an indefinite number of times and can be improved and transformed through the centuries while safeguarding all that is best and most worthwhile of what has been established earlier.

POETRY

And now we must come to a question which has long lain in wait for us. Will the machine translate poetry? To this there is only one possible reply—why not? All of us have done it in our schooldays, when neither our Latin syntax, nor our grammar, nor our vocabulary, nor our sense of rhythm, nor our skill in rhyming could rival those of the electronic machines of tomorrow. Do not let us ask the machine to do more than a minimum, but let us see what this minimum may be, how we can, if possible, improve upon it, and what lessons it can provide for the future.

The task of the translator becomes progressively more complicated and sophisticated as the text for translation grows further removed from straight description or narrative, as its vocabulary becomes more connotative and less denotative, and as extra-linguistic elements, such as the elements of situation in dialogue of novels or plays, take precedence over those strictly linguistic markers sought in the sentence by the programmers of automatic translations. In dialogue, in poetry, in “stream of consciousness” writing, in everything which suggests a momentary individual representation rather than a Cartesian expression of a clearly defined concrete or abstract reality, the task of the translating machine will become extremely difficult. Not that there is any difference in kind between this language and the language of scientists; but the search for lexical equivalents between two languages becomes more problematical and depends on a greater number of factors, some of them extralinguistic. The private understanding between author and reader of a scientific text does not necessarily exist between the reader and the author of a poem. The more the choice of words becomes an individual matter instead of being dictated by the constraints of a scientific discipline, the greater will be the number of sub-routines required by the machine or by the translator for lexical research, and the less likely it is that such research will prove economic or even possible. How can the machine succeed in a domain where the magic of sound and rhythm, of extraneous semantic evocation are the imponderable guides of the sensitive translator?

Between metalanguage and pure poetry, from the clear and distinct expression of a scientific representation to the synthetic expression of the vibrations of the poet’s ego at the centre of his individual universe, there exists a whole vast range of untranslatables. All translation is an approximation, because language alone is translated while metalanguages require no translation. If we dare to reply “why not?”, it is because from the Cartesian absolute of metalanguage to the mystic absolute of pure poetry, there are differences not of kind but only of degree.

Poet or geometrician, the true writer gives to language both its full connotative and musical harmonics and its full denotative value. Is it not possible that by tackling boldly the difficulties of poetic translation, we may hit upon the solution to some of the more profound problems of scientific or narrative translation? It would be foolish to assert that the machine will translate poetry as successfully as a handful of poets have done, but it would be worse than folly, once the instrument has been forged, not to make use of it to take the measure of its failures as well as of its successes, and to find out where and why it fails.

STUDIES IN POETICAL SEMANTICS

Such an analysis would be based on many aspects of language which it has not been possible to discuss in the present study. Together with work on written language and its translation we have mentioned briefly the existence of research into the phonic aspects of language. Sooner or later the study of written language and that of spoken language will combine not merely for oral control of the movements of elevators, for automatic simultaneous interpretation and stenotyping, but also for the study of the rhythms and sounds of poetry and of prose. Electronic recording will make it possible to conserve and to classify sounds and rhythms, and to submit them to exact and objective study of a type impossible for the critic working with only intuition or card indexes to guide him. In the same way studies of comparative semantics, indispensable for the constitution of electronic bilingual dictionaries for machine translation, will be directed beyond the bounds of utilitarian requirements towards the semantic analysis of poetry, which will take into account both suggested and expressed meaning. Matila Ghyka, noting Mallarmé's poetic predilection for the words *azur, vierge, or, cristal, glacier,* has sketched the outline of a semantic study of the sound and "shape" of words, showing the way to a purely disinterested research to which the new machine methods can contribute. Thorough studies of comparative semantics will make it possible to determine to what extent it is possible to find equivalents, in other languages, for the connotative value of words, so vital a factor in poetry. Already Reifler's work contains interesting suggestions in this direction. Automatic translation workshops should be so managed that they can put at the disposal of literary research part of the time of their machines and of the experience acquired by their personnel in thorough and penetrating analysis of all aspects of language.

Electronic machines have composed lyrics and verses, have invented rhymes and composed music. Will there one day emerge from the sorry mechanical monotony of these attempts, as sometimes from anthologies of anonymous poets of bygone ages, lines which at least evoke, let alone create, the fleeting thrill of human emotion? This is at least as statistically probable as the recreation of the works of Shakespeare by a monkey blindly typing through eternity on the keys of the typewriter. Will man be able to lead the machine beyond this random search for beauty, and so direct it that he can speed up its creative course, thus suspending, at least for a moment, the rule of chance? To attempt to translate poetry by machine, after full analysis of the constitutive elements of poetry, is a more alluring proposition than to teach a robot to make rhyming couplets. It is a proposal that should attract lovers of poetry as well as iconoclasts—all those, in fact, who wish to penetrate the secrets of verbal creativeness.

LITERARY ANALYSIS

In fact the problems of automatic translation is here only one of the many aspects of the application of electronic machines to literary analysis. Linguistic data systematically registered in magnetic memories or on punched cards or tape can be subjected to a number of analyses in the same way as can scientific problems or questions of management in complex business enterprises. If the translation of scientific texts represents a utilitarian aspect of this new science—the analysis of discourse with methods offered by electronics—it is also true that computers make it possible to submit *all* the elements of spoken and written discourse to a systematic study impossible with the individual, manual methods of yesterday. Father Roberto Busa has demonstrated this by his studies of the work of St Thomas Aquinas and the *Dead Sea Scrolls.* He has perfected a method of making, within a relatively short period, a concordance of the *Summa Theologica* and an index of the *Dead Sea Scrolls.* In the latter case the mechanical analysis programme has permitted the reconstitution of words missing in the manuscript, on the basis of studies of the frequency of certain word groups. The application of similar techniques to literary, juridical and scientific studies is still in its infancy, but an attempt is already in progress to extend it to automatic summary records and abstracting.

In a short chapter in their book, [7] Booth, Brandwood and Cleave have summarized the various aspects of contextual and structural analysis made possible by the use of electronic computers: frequency counts on the lines of those of Estoup and Zipf, but employing more rapid and more reliable techniques; biblical and other concordances; the constitution of a dictionary of syntactic structures. They deal at greater length with stylistic analysis, the chronology of the works of Plato and the mechanical study of rhythms and syntax in the *Dialogues.* Their book provides us with glimpses of a whole new technique shortly to be at the disposal of literary research, which cannot afford to neglect these new methods.

How much time will such analyses require in detailed preparation for the high-speed work of the machine? This, Booth replies, will depend on the number of persons working on the problem. It is no longer a case of a work of laborious scholarship undertaken by one man at the beginning of a lifetime of patient work: there must be a new division of labour, with a hierarchy for the formulation of exact rules to be strictly applied by all. A hierarchic division of labour on these lines is already apparent in the team of Father Busa at the Aloysianum at Gallarate, where leadership is in the hands of the inventor of the research while the tasks of execution are spread among less ingenious technicians. Thus, in order to make use of modern techniques, literary research will have to become collective, as scientific laboratory research already is. This evolution must be fully comprehended and controlled so that we do not fail to safeguard the essential: respect for and knowledge of the creative genius of man, the secret of which both translator and analyst are endeavouring to fathom. Machines bring us the means to know and to understand writers better, in an age when there is a greater need than ever before for the human community to affirm the right of all men to culture and to knowledge, and to see to it that every nation and every individual has the means to make this right a reality.

SPEEDING UP CULTURAL EXCHANGE

Chance alone has not decided Soviet scientists to work on the elaboration of automatic translation programmes into Russian from Arabic, Burmese, Chinese, Hindi, Indonesian and Vietnamese. Nor is the importance attached to Asian languages by Soviet linguists entirely political. Newly independent nations need to be able to read in their own language the works of other countries. They would also like to see their own literature translated into other tongues. The automatic translation programme into Russian represents the first step in the process of exploring the linguistic relationships between languages which will end in two-way translation programmes. As we observed in Chapter I, automatic translation can accelerate the contacts of these young nations with other peoples, helping them to affirm their personalities and to safeguard their own cultural heritage, enriching it at the same time by contact with other cultures. This would seem to be the only way of crossing before it is too late, without compromising the originality and diversity of cultures, the barriers raised between peoples by linguistic difference. In a world where cosmopolitan currents establish themselves thanks to swiftness of communication, and where a small number of languages might, for economic and strategic reasons, come to impose their hegemony over the whole world, to multiply authentic translations is one way of defending the profound originality of national cultures, against the tendency towards standardization brought about by more and more uniform techniques and by the world-wide spread of a universal technical terminology.

WHAT REMAINS TO BE DONE?

The translation machine, together with revolutionary new techniques in linguistic analysis, is now on our doorstep. In order to set it to work, it remains to complete the exploration of linguistic data by means of comparative lexical and structural analyses, first bilateral, then multilingual. If, in these pages, we have devoted more space to language than to machines, it is because the prime necessity is the adaptation of linguistic studies to the new techniques. It is impossible to emphasize this point too strongly: machines capable of translation already exist—it remains for men to learn how to make use of them.

To inventory words and meanings, to undertake statistical studies of semantic frequency, to catalogue inflexions and their grammatical functions, to analyse word order and its exact significance or value, to list types of structure and their meaning—these first tasks can be greatly facilitated by the use of tabulators and computers. Simultaneously, or perhaps subsequently, it will be necessary to plan, organize and keep up to date a large electronic dictionary fully adapted to the potentialities of existing machines, and constantly to improve the operation of this dictionary in the light of the evolution of machine techniques. It will be useless to await the perfect machine before setting to work. Nature did not await the human brain before creating nervous tissue. As soon as a dictionary designed and compiled for the translating machine is registered on magnetic tape, it will be an easy matter to transfer it on to any alternative form of memory offering greater speed of access or higher capacity. Technically, the go-ahead signal has been given and we can count on the new techniques to facilitate progress in the tasks they have opened up to us.

The third task, which can be tackled at the same time as the first two, will be the construction of a machine specially adapted to translation needs. Booth has somewhat optimistically estimated its cost as between £50,000 and £100,000. Such a machine would be of no avail without programmes, and without it the best programmes would serve no useful purpose. Meanwhile, it would be a waste of more powerful machines, capable of more complex operations, to use them continually to perform translations demanding no operation more complex than addition and subtraction. Thus programming can begin before the ideal machine is available, but should be undertaken only in close collaboration with the technicians who know existing machines and are able to design new ones fully adapted to translation.

No attempt has been made here to conceal or omit all that still remains to be done. We have tried to give a synthesis of recent work without wearying the reader with too many technical details. A great many problems in fact remain unsolved. But the way is now open and one solution often leads to another. An attempt has been made to explore the complexity of linguistic data, and it has already been established, for example, that the translation of a twenty-word sentence may require as many as

10,000 logical machine operations. If the translation of more complex sentences is to be performed rapidly enough to be economically interesting, we shall have to discover more expeditious methods, making it possible to reduce such operations to a minimum and economize some millionths of a second per word. The exploration of linguistic structures will have to be pursued to the very end, so that we may discover whether it is really possible to translate not word-by-word but clause-by-clause, as anticipated by the structuralists. We shall have to solve the problems of self-programming, so that the machine can choose for itself the most effective programme for a given structure.

If much remains to be done in the field of simplification and rationalization of programmes, the same is true of the acceleration of machine input methods. Tape-punching is slow and relatively costly, particularly for Cyrillic, Arabic and Asian scripts and for ideographic languages. Direct reading by the machine, with automatic coding of printed text, without human intervention, would be the ideal solution, and work is already in progress to this end. These are but a few examples of the type of problem with which it has been impossible to deal here since such questions are subsidiary to the fundamental analysis of language for automatic translation. Finally, it must be added that if the perceptron, a new machine, based on the ideas of Ross Ashby, fulfils its early promise and can be trained to recognize patterns, it should provide the solution to the remaining problems of syntax and structure.

While a great deal remains to be done, it can be stated without hesitation that the essential has already been accomplished. Before broaching the problem and taking stock of it, men had to free themselves of taboos; that having been done, the rest is a matter of technique only. Booth's bold thinking, Reifler's patient linguistic ingenuity, Panov's scientific dialectics and empiricism, have won acceptance for a new attitude towards the study of language—an attitude which, while respecting the individual qualities of a spoken or written text, is nevertheless fully determined to explore these qualities and as far as possible to imitate them in another language. The translating machine recalls to mind a very simple tool, the use of which has long since ceased to be considered sacrilegious: the pantograph, with which a workman can copy, in the material of his choice, the marble Venus of Milo, without disrespect for the inspiration and genius of the unknown sculptor. Translating machines will soon take their place beside gramophone records and colour reproductions in the first rank of modern techniques for the spread of culture and of science.

Postscript to the English Edition

THIS book was completed in its original French version by the end of December 1958 and published in May 1959. In June of this year the author attended the Unesco conference on information processing, which enabled machine translation specialists to meet and exchange views on the state and prospects of their work. With few notable exceptions most of the schools of research mentioned in the foregoing pages were represented.

The conference provided an opportunity to have a mechanical translation made, without previous trial or preparation, of a foreword specially written in French for this English edition so that it might be so translated. No choice of programme or of machine was possible. It so happened that Mr A. F. R. Brown of Georgetown University had brought to Paris his recorded French-to-English translation programme and vocabulary, designed for the translation of texts on chemistry and nuclear energy; the vocabulary is of some 4,000 words and the programme operates at the speed of 5,000 to 10,000 words per hour, using an IBM 704 computer available in Paris at the I.B.M. headquarters.

The text of the foreword was given in French typescript to Mr Brown at 5.30 p.m. on 19th June 1959. He proceeded to the I.B.M. headquarters where he keypunched it; the text as entered into the machine is shown in Figure 1. The figure “1” following a letter means “acute accent” while the figures “2” and “3” conventionally designate the grave and circumflex; “$ FIG” means that the signals following it are figures and not letters of the alphabet, “$PAR” meaning “new paragraph”.

The translation, produced and handed over by 6.30, is also reproduced in Figure 5. From it the reader can deduce that the words have been analysed into stems and affixes in such a way that the mark of the acute accent in “spécialisé » is retained in English in the translation “specialised”. This seems due to the literal transcription of stems which are alike in English and French. Words not recorded in the memory of the machine because they are not relevant to the vocabulary of chemistry and nuclear physics, such as "calculatrices", "traductions", "langues", etc., appear unchanged in the English output. One word, "encore", seems to have caused some technical hitch, coming out as it did as "000017".

This programme is clearly still inadequate for the proper translation of present participles used as adjectives, such as "satisfaisant" and "exigeant", which have been treated as verbs and not subjected to the rule for changing word order—unlike "spectaculaire" which has been correctly translated and placed.

Similarly, prepositions require further programming—e.g. to avoid such renderings as "to present at the reader" ; past participles (définies) and nouns having the same forms as past participles (découvertes) also, as do pronominal verbs (s'améliorer) and some adjectives (large Britain).

The machine does not work without humour. The programme being originally designed for chemistry, words such as "brom-ure", "iod-ure", "carb-ure" are stripped of their chemical affix and rendered by "bromide", "iodide" and "carbide". Hence—and that is a fault in programming—the feminine adjective form "future" has been wrongly stripped of a false affix and a new chemical has been invented by the machine, "futide"!

One alteration was suggested by Mr Brown before keypunching: the addition of "en" before "croissant." But no attempt was made by the author to simplify style or to avoid idioms such as "se tenir au courant" or "mise au point", mechanically translated by "hold at the current" and "put at the point".

Two Russian-language versions of the same *Avant-Propos* were made by Mr Michael Corbe: one of them follows the French original almost word for word; the other is on the contrary free and easy. Mr Corbe left both versions with Dr Don Swanson of the Ramo-Wooldridge Corporation at Los Angeles, where both versions were in their turn translated into English by means of Dr Swanson's experimental programme for the translation of Russian physics texts. This programme is stated to be capable of producing considerably better than "word-for-word" translation. It is of course subject to the same lexical shortcomings as all such programmes, and indeed as any human being, in that it cannot translate words which it has never encountered before, and which are therefore not recorded in its memory. On the other hand it is well-equipped for syntaxical analysis, on which Dr Swanson's team has concentrated, and it has an efficient method for stripping off and interpreting in English the flexional endings of Russian.

The Appendix contains, set in four columns, four of the texts involved in this multiple experiment: column I shows in sequence, numbered vertically, each word or coherent group of words of the original French text. Facing each French horizontal line, the reader will find in column II, presented in the same manner, Mr Brown's machine translation into English, together with (in italics) those indispensable editorial amendments where French words had been used by the machine instead of their English equivalents. Here the wide lexical similarity of English and French through their common Latin stock of words, has undoubtedly facilitated the task of the machine and of the post-editor.

Column III shows, similarly laid out, Michael Corbe's "human" word-for-word translation. Finally column IV gives Dr Swanson's machine-made English translation from that particular Russian text; it has been so arranged as to show in italics those editorial improvements introduced at Ramo-Wooldridge Corporation in pencil on the actual machine output, as part of the cumulative process of lexical amelioration of the programme. It will be seen that those lexical editorial changes are quite numerous whereas the grammatical structure has on the whole been very accurately rendered by the machine. Moreover, for each Russian polysemantic word the machine has given alternative translations; the one selected by the post-editor has been given first. Such editorial choices are later recorded in the memory with contextual information as part of the cumulative improvement process. A similar procedure is applied to idioms, such as *deržat'sja v kurse,* the correct rendering of which will now have been recorded for future use in the Ramo-Wooldridge programme.

Careful examination of those four columns in the Appendix will it is hoped give the reader a fairly precise idea of the present state of machine translation and of the practical means by which it is being improved.

The Ramo-Wooldridge version of Mr Corbe's free and easy Russian translation has not been added here as a further illustration. It differs from the translation shown in column IV by its better flow of English phrase—while it suffers from the same lexical shortcomings.

The part of the conference and the symposium devoted to mechanical translation permits certain conclusions which strengthen the main thesis of this book. Machine translation is now not only possible, it is actually being carried out, but not as a finished product and mainly in an experimental and fact-finding spirit. Its end-products are and will be many, and perhaps at this stage the least important is translation itself. The method of gradual accumulation of carefully checked data, using the machine as a means of objective analysis of language, is now consecrated in the work of Harper and Hays at Rand Corporation, of Oettinger and his team at Harvard; it appears to be gaining ground even among the research groups with a more theoretical approach. This method makes the fullest use of the electronic computer and its ancillary machines, sorters, tabulators, etc., to subject the data recorded about words in a text and its translation, to successive analyses from various points of view. Language data are indeed processed not only with translation in mind but with the aim of obtaining the widest and deepest penetration of such facts as the relationships between words. Harper and Hays have in particular presented a method of analysis of structures based on the dependency and precedence relationship between words which promises considerable simplification in structural analysis by machine.

It was unfortunate that neither Miss Bel'skaja, who had submitted a remarkable paper, nor Professor Panov who had prepared the survey on the present state of machine translation, could attend the conference owing to illness. Russian work in this field was ably represented by Miss Kulagina, whose approach is based on mathematical theory—this at a time when the more empirical approach of Panov and Bel'skaja appears to be winning support even from the more theoretically-minded Western research teams. It is to be hoped that the original views expressed by Bel'skaja on the feasibility of translating poetry by machine will be developed and clarified, and also that she will be able to give wider dissemination to her sub-routines for the analysis of English sentences.

From these meetings it clearly emerges that M.T. has reached a stage where theories must temporarily recede into the background while practical laboratory work and machine processing of long consecutive texts is indispensable to further research and to the development both of practical programmes and of new working hypotheses, based on the study of numerous facts rather than on intuitive preconception. This perhaps did not require proof, but it will not be amiss, for the future of research in language-data processing, that this international conference brought it out in the full light.

The critical attitude of M.T. specialists towards traditional grammars was moderate and tactful; it was felt that they need improving and completing rather than scrapping, and that to start from them and work on their improvement is better than to start from scratch.

Another fact which struck members is the international character of this type of research, and the need for close co-operation between national centres conducting language studies. More instructive perhaps than the public meetings were the small informal gatherings in which specialists—most of them very young—exchanged information about details of their methods of analysis and of recording facts in the memories of the computer and of arranging programme routines.

The general conclusion is one of optimism. Machine translations today are still very imperfect. But the way to perfecting them is clear. The field is attracting talented people in increasing numbers. One of the major problems is to produce programmes which do not ramify into excessively time-consuming sub-routines while solving most of the problems of sentence structure if not of polysemy. And the growing experience of programmers points to man's ability to observe the behaviour of the machine and give it a chance to solve simply problems which at first baffle the mind because we have not yet learned to state them simply. We can trust the machine to teach us precisely that, because its fundamental methods are simple.

Bibliographical Notes

[1] Andreev, N. D., "Mechanical Translation and the Problem of an Intermediary Language" (in Russian) in *V.J.,* 6 (5). (See below No. 30.)

[2] Ashby, William Ross, *Design for a Brain,* Chapman & Hall, London, 1952.

[3] Ashby, William Ross, *An Introduction to Cybernetics,* Chapman & Hall, London, 1956.

[4] Belevitch, Vitold, *Langage des machines et langage humain,* Hermann, Paris, 1956.

[5] Bel'skaja, I. K., "Machine Translation of Languages," in *Research* 10 (10), October 1957, Butterworths, London. (See also No. 29 below.)

[6] Booth, A. D. and K. H. V., *Automatic Digital Calculators,* Butterworths, London, 1953.

[7] Booth, A. D., Brandwood, L., and Cleave, J., *Mechanical Resolution of Linguistic Problems,* Butterworths, London, 1958.

[8] Booth, K. H. V., *Programming for an Automatic Digital Calculator,* Butterworths, London, 1957.

[9] Cary, Edmond, *La traduction dans le monde moderne,* Georg, Geneva, 1956.

[10] Cherry, Colin, *On Human Communication,* Technology Press of the MIT and John Wiley & Sons, New York, Chapman & Hall, London, 1957.

[11] Communications Research Centre, University College, London, *Aspects of Translation,* Secker & Warburg, London, 1958.

[12] Delavenay, E. and K. M., *Bibliography of Mechanical Translation,* Mouton & Co., The Hague (in preparation).

[13] Fries, Ch. C., *The Structure of English, an Introduction to the Construction of English Sentences,* Harcourt, Brace & Company, New York, 1952.

[14] Guilbaud, G. Th., "La cybernétique." *Collection Que sais-je?* No. 638, Presses Universitaires de France, Paris, 1957.

[15] Hays, David G., Harper, K. E., and others, RAND research memoranda: *M.T. Studies 1-9,* RAND Corporation, Santa Monica, California. (See also No. 29 below.)

[16] Liebesny, F., "Economics of Translation," in *ASLIB Proceedings,* 10 (5), London, May 1958.

[17] Locke, W. J., and Booth, A. D., and others, *Machine Translation of Languages.* Technology Press of the MIT & John Wiley & Sons, New York, Chapman & Hall, London, 1955.

[18] Kulagina, O. S., and Mel'čuk, I. A., "Machine Translation from French into Russian" (in Russian) in *V.J.,* 5 (5). (See below No. 30.)

[19] Mološnaja, K. T., "Some Syntactic Problems in Machine Translation from English into Russian," (in Russian) in *V.J.,* 6 (4). (See below No. 30.)

[20] Mounin, Georges, *Les Belles Infidèles,* Cahiers du Sud, 1955.

[21] Mounin, Georges, "Pseudo-langues, interlangues et metalangues," in *Babel,* 4 (2), June 1958.

[22] *M.T. = Mechanical Translation,* devoted to the translation of languages with the aid of machines, Massachusetts Institute of Technology, 1954-.

[23] Muhin, I. S., *Opyty avtomatičeskogo perevoda na elektronnoj vyčislitel'noj mašine B.E.S.M.,* Academy of Sciences of the U.S.S.R., Moscow, 1956.

[24] Murphy, John S., *Basics of Digital Computers,* Rider, New York, 1958.

[25] Panov, Dmitrij J., *Avtomatičeskij perevod,* 2nd edition, revised and augmented, Academy of Sciences of the U.S.S.R., Moscow, 1958.

[26] Panov, Dmitrij J., *Concerning the Problem of Machine Translation of Languages,* Academy of Sciences of the U.S.S.R., Moscow, 1956 (in English).

[27] Sestier, A., *Les calculateurs numériques et leurs applications,* Hommes et techniques, Neuilly-sur-Seine, 1958.

[28] *Tezisy konferencij po mašinnomu perevodu,* 15th-21st May 1958, Ministry of Higher Education, Moscow, 1958.

[28a] *Tezisy soveščanija po matematičeskoj lingvistike,* 15th-21st April, 1959, Ministry of Higher Education, Leningrad, 1959.

[29] Unesco: *Proceedings of the International Conference on Information Processing.* Unesco & Oldenbourg, Munich; Butterworth, London; Dunod, Paris (in preparation).

[30] *Voprosy Jazykoznanija,* Academy of Sciences of the U.S.S.R., Moscow, 1955 and after.

Glossary

ALGORITHM or Algorism: in general, the art of calculating with any species of notation; in particular the word is used by computer programmers to designate the numerical or algebraic notations which express a given sequence of computer operations, define a programme or routine conceived to solve a given type of problem.

BINARY CODE: A binary system of numbers or other marks (e.g. electronic pulses, etc.), used to represent either decimal digits or letters of the alphabet. See *binary system* below for the binary coding of decimal digits.

BINARY SYSTEM: A system of counting using two as the radix, or base, whereas the more familiar decimal system uses ten as the base. Only two characters or symbols are used, 0 and 1, or + and —, or, in electronic circuits, pulse and no pulse, whereas the decimal system uses ten characters. The following table shows the conversion from decimal symbols to binary:

Decimal	*Binary*	*Decimal*	*Binary*
0	0	5	101
1	1	6	110
2	10	7	111
3	11	8	1000
4	100	9	1001

Because electronic pulses and magnetic states have binary form (on or off, pulse or no pulse, + or —) they lend themselves easily to the recording of data in computers.

COMPLEMENTARY SIGNALIZATION: A device consisting of adding conventional signals to existing alphabetical signs, first developed by Professor Reifler to "pre-edit" sentences prior to machine-translation. This system was abandoned with the progress of computers and of language analysis for M.T., as it was found that where the human mind unconsciously recognized signals in written sentences, the machine can in most cases be programmed so as to recognize them too.

CRYPTOGRAPHY: The act or art of writing in secret characters or cipher: the science or techniques of cipher.

CYBERNETICS: A word derived from Greek *kybernētikē,* the art of steering a ship, helmsmanship ; used by Ampère (1834) to designate the study of means of government; then by N. Wiener in *Cybernetics, or Control and Communication in the Animal and in the Machine,* Hermann, Paris 1948. Wiener chose the word under the influence of the Watts "governor" on the steam engine, one of the earliest feedback controls ever invented. Cybernetics is used with precision to mean the science of control mechanisms, and loosely to designate the theory and practice of automata and calculating machines, "thinking" machines, etc. See Bibliography, J. Th. Guilbaud, for a sober appraisal of this new science.

DESINENCE: Termination, ending of a word—more properly the inflected ending of a word.

DIACRITICAL SIGNS: Distinguishing signs, marks, points or other signs, attached to a letter or symbol to distinguish it from another of similar forms: e.g. accents in French.

DIGIT: (Latin *digitus,* finger): each of the numerals below 10 in decimal counting, 0 to 1 in binary (q.v.) counting.

DIGITAL COMPUTER: See DIGIT. As opposed to the *analog computer,* which simulates the problems it is asked to solve, the *digital* computer, derived from Pascal's arithmetic machine and from the desk calculator, works out numerical solutions to problems, by calculations made with and on digits.

HOMOGRAPH: One or two or more words identical in spelling, but of different derivation or meaning: e.g. French *la* route, suivez-*la*, etc.

HOMONYM: A word having the same pronunciation as another, but differing from it in origin, meaning and possibly spelling.

Hence, HOMONYMY. Homophones are homonymous in sound alone, homographs are homonymous in spelling.

HOMOPHONE: A letter or word having the same sound as another, but differing from it in meaning and/or spelling. E.g. French ou, où, houe; English: Rome, roam; bare, bear.

INFORMATION THEORY: An approach to the study of messages transmitted in a language, based on the mathematical theory of communication, as developed by communication engineers in their search for economy and efficiency in transmission of messages. The statistical study of language in so-called information theory bears mainly on the frequency of reference of graphemes and phonemes, but can be developed in various directions more directly useful to applied linguistics and M.T.

INPUT ORGAN: Any organ of a computer through which data are fed into it—e.g. a punched-card reader, tape-reader, photo-electric reader, etc.

INVARIANT: An invariable quantity—a term from the vocabulary of modern mathematics. A semantic invariant is a constant semantic fact which is found in languages having historical or other connections, such as the evolution, in accordance with certain laws, which would seem to be universal, of the meanings of words first designating an object but later acquiring an additional meaning. For instance French "mouton", sheep, "moutonner" said of a cloudy sky where the small clouds suggest a flock of sheep: a similar semantic evolution is found in Chinese.

LEXICAL CONTENT: The word-content of a language or a sentence, book, etc.—its vocabulary.

LEXIS: A greek word meaning word, phrase, diction. Used here, and by M.T. linguists, to designate the words of a language, contained in its dictionary or lexicon, as opposed to the morphology and syntax of that language.

MACROSCOPIC, MICROSCOPIC: Greek *makrós,* great, *micrós,* small As opposed to microscopic study, which rivets its attention on infinitely small details, macroscopic study concentrates on large-scale aspects of phenomena—for instance macroscopic linguistics bears on very general statistical rules of language (e.g. Zipf's law) rather than on individual aspects of language or speech. Information theory (q.v.) has so far studied mainly macroscopic aspects of language.

MICROSECOND : One millionth of a second.

MILLISECOND: One thousandth of a second.

MORPHEME : A form which cannot be analysed into smaller forms, together with its corresponding meaning.

MORPHOLOGY: That branch of language study which deals with the functions of inflexions and derivational forms, hence MORPHOLOGICAL, pertaining to —.

M.T.: Machine translation, or mechanical translation.

PHONEMICALLY: A phoneme being a group of variants of a speech sound (e.g. the *e* sound in *get, tell, say, any, send)*—phonemic means "of the nature of a phoneme"—also "significant, distinctive" (of sounds). Hence "phonemical", "phonemically".

PHONETOGRAPH: An instrument designed to record the sounds of speech in the form of typewritten sequences of letters of the alphabet.

POLYVALENT: A word from the vocabulary of chemistry, where it means "having multiple valence". By extension: potentially capable of fulfilling several functions, playing several roles.

POLYSEMY: Multiplicity of meaning.

POLYSEMANTIC: (A word) having several meanings. Polysemantic dictionary: a dictionary of words which have the common characteristic of each having several meanings.

SEMANTICS (also semasiology): The branch of philology which deals with meanings. Used here in contrast with syntax, morphology and even with lexis, which is the set of words of a language, as opposed to the various meanings of a word, studied by semantics.

SEMANTIC INVARIANT: See INVARIANT.

SEMANTIC UNIT: A unit of meaning, as opposed to units of vocabulary, or to phonemes or morphemes. In *amavi, am-* expresses “love” and *"avi"* the 1st person in the preterite: both have semantic value, each is a unit. In *je n'ai qu'un livre, ne ... qu'* is a semantic unit made up of two words.

SEMANTIC VALUES: The word “dog” has different semantic value when it designates the animal, or a “fire-dog”.

SEME : From Greek *sēmeion,* a mark, sign, a unit of meaning.

STATISTICAL SEMANTICS: Statistical study of meanings of words and their frequency and order of recurrence.

STRUCTURAL LINGUISTICS: A form of the scientific study of language which concentrates on structures or patterns (“he—the—a—” is such a pattern, which can be filled in as “he gave the car a push”, “he found the boy a drink”, etc.). Some schools of structural linguistics tend to disregard *meaning* as unessential to the study of structure; all emphasize the paramount importance of patterns in the development of language and in its teaching.

SYNTACTIC: Pertaining to SYNTAX q.v. *Syntactic value,* significance in terms of relations between words, of their syntactic link (position, preposition, etc.).

SYNTAGMA: Arrangement of units in a syntactic construct, such as “actor+action+goal” (the dog bit the man) .

SYNTAX: That branch of linguistics which deals with the arrangement of syntagmata. *Syntactic analysis* for M.T. is the analytical study of word-arrangement, with a view to programming translation work in such a way that sentence-for-sentence translation will be possible, as opposed to word-for-word.

TECTONIC : Structural, pertaining to the structure of the sentence or group of words.

AVANT - PROPOS.

PRE1SENTER AU LECTEUR QUI N EST SPE1CIALISE1 NI
DANS L E1TUDE DE LA LING UISTIQUE NI DANS LA
CONNAISSANCE DES CALCULATRICES E1LECTRONIQUES ,
LES PROBLE2MES ACTUELS DE LA TRADUCTION
AUTOMATIQUE DES LANGUES , TEL EST LE BUT DANS
LEQUEL CE LIVRE A E1TE1 CONCU .
NOMBREUSES SONT LES DIFFICULTE1S QUI SE DRESSENT
ENCORE SUR LE CHEMIN AVANT QU UNE TRADUCTION
SATISFAISANTE POUR UN LECTEUR UN PEU EXIGEANT
PUISSE SORTIR D UNE MACHINE ,.. DE GRANDS
PROGRE2S ONT E1TE1 RE1 ALISE1S DANS L ANALYSE
DES LANGUES , ET LES PRINCIPALES ElTAPES DE
L E1TUDE DU LANG AGE EN VUE DE LA TRADUCTION
AUTOMATIQVE SONT MAINTENANT DE1FINIES .
LA RECHERCHE A PROGRESSEl DE FAÇON SPECTACULAIRE
DEPUIS $FIG 1955 ,.. DES TRADUCTIONS UTILES SONT
FAITES PAR DES MACHINES ET LEUR NOMBRE IRA EN
CROISSANT , LEUR QUALITEl S AME1LIORERA
CONSTAMMENT .
MAIS CERTAINES DE1COUVERTES SONT NE1CESSAIRES
POUR QUE CETTE RECHERCHE ENTRE BIENTO3T DANS UNE
NOUVELLE PHASE , CELLE DE L AUTOMATISATION. A2
$FIG 98 OU $FIG 99PCT .
$PAR LE LECTEUR QUI SOUHAITE SE TENIR AU COURANT
ET SUIVRE CETTE FUTURE E1TAPE DU PROGRE2S
SCIENTIFIQUE TROUVERA ICI UNE MISE AU POINT DE L
E1TAT ACTUEL DES TRAVAUX TEL QU IL RESSORT DES
OUVRAGES ET ARTICLES PARUS DE PUIS $FIG 1955 AUX
E1TATS - UNIS , EN GRANDE - BRETAGNE ET DANS L
UNION SOVIE1TIQUE .

FIG. 5. A SPECIMEN OF MACHINE TRANSLATION

(a) **A Foreword to this book, as typed out in its original French in the course of its mechanical translation on I.B.M. 784 computer. This Foreword was written for the sole purpose of being so translated. See page 119 for an explanation of figures in words and other conventional symbols.**

BEFORE - REMARK .

TO PRESENT AT THE READER WHICH IS SPECIALISED NEITHER IN THE STUDY OF THE LINGUISTIC NOR IN THE KNOWLEDGE OF THE ELECTRONIC CALCULATRICES , THE PRESENT PROBLEMS OF THE AUTOMATIC TRADUCTION OF THE LANGUES , SUCH IS THE AIM IN WHICH THIS BOOK HAS BEEN CONCEIVED .
NUMEROUS ARE THE DIFFICULTIES WHICH SET UP THEMSELVES 0000170N THE PATH BEFORE A TRADUCTION SATISFYING FOR A READER A LITTLE REQUIRING CAN EXIT FROM A MACHINE /COLON/ LARGE ADVANCES HAVE BEEN REALISED IN THE ANALYSIS OF THE LANGUES , AND THE PRINCIPAL STEPS OF THE STUDY OF THE LANGUAGE IN VIEW OF THE AUTOMATIC TRADUCTION ARE NOW DEFINITE .
THE RESEARCH HAS PROGRESSED IN A SPECTACULAR MANNER SINCE 1955 /COLON/, USEFUL TRADUCTIONS ARE DONE BY MACHINES AND THEIR NUMBER WILL INCREASE CONTINUOUSLY , THEIR QUALITY WILL IMPROVE ITSELF CONSTANTLY .
BUT CERTAINES DISCOVERED ARE NECESSARY FOR THIS RESEARCH TO ENTER SOON A NEW PHASE , THAT OF THE AUTOMATISATION AT 98 OR 99PCT .
THE READER WHICH SOUHAITE HOLD AT THE CURRENT AND FOLLOW THIS FUTIDE STEP OF THE ADVANCE SCIENTIFIC FIND HERE A PUT AT THE POINT OF THE STATE PRESENT OF THE WORK SUCH THAT IT BE EVIDENT FROM THE WORK AND ARTICLE PARU SINCE 1955 AT THE STATE UNITE , IN LARGE - BRITAIN AND IN THE UNION SOVIE1TIC .

(b) Reproduction of the actual machine-translation of the same Foreword, as typed out by the I.B.M. 704 computer in Paris on 19th June 1959. The French-to-English translation programme used was conceived and designed by Mr A. F. R. Brown of Georgetown University for the translation of texts on chemistry and nuclear energy. A fuller explanation will be found on page 119.

See also Appendix for a comparison of this translation with a machine-translation of the same text from a Russian version.

APPENDIX

Two Machine Translations of the Same Preface to this Book

The Comparative Tables below show:
I. French original of a Preface written for this book; II. Machine translation of this original from French into English by means of a programme devised by Mr A. F. R. Brown of Georgetown University, on an I.B.M. 704 computer; translation made in Paris on a computer of this type; III. Word-for-word Russian version of the same French Preface, prepared by Mr Michael Corbe; IV. English version of this Russian text, machine-made at the Ramo-Wooldridge Corporation, Los Angeles, on an IBM 704 computer, under the direction of Dr Don Swanson.

Each text is presented vertically, one word or coherent group of words at a time on one line. Comparison can thus be made between any two or more vertical columns to see what happened to a given word.

KEY to column II:

ROMAN TYPE: machine output.

ITALICS : English translations, supplied by the author, of those words for which the machine did not possess a translation in its memory.

KEY to column IV:

ROMAN TYPE: machine output in English, as accepted by Ramo-Wooldridge post-editor.

ROMAN TYPE in square brackets: alternative machine output in English, rejected by post-editor.

SMALL CAPITALS: machine output in Russian, i.e. words which the machine could not translate because they were not in its memory.

ITALICS: English words supplied by the post-editor as part of the process of cumulative improvement of his programme.

	I *Author's Original French*	II *Brown's Translation, I.B.M. 704*	III *Corbe's Translation, manual*	IV *Swanson's Translation, I.B.M. 704*
1	Avant-propos.	Before-remark.	PREDSILOVIE.	Preface.
2	Présenter	To preesnt	PREDSTAVIT'	To present [to represent]

	I *Author's Original French*	II *Brown's Translation, I.B.M.. 704*	III *Corbe's Translation, manual*	IV *Swanson's I.B.M. 704*
3	au lecteur	at the reader	ČITATELJU	to the reader
4	qui	which	KOTORYJ	which *(that)*
5	n'		NE	is [there is, there are]
6	est	is	EST'	
7.	spécialisé	specialised	SPECIALIZIROVAN	SPECIALIZIROVAN *(specialised)*
8	ni	neither	NI	neither [nor, not]
9	dans	in	V	in
10	l'étude	the study	IZUČENIJ	the study
11	de la linguistique	of the linguistic	LINGVISTIKI	of the LINGVISTIKI *(of linguistics)*
12	ni	nor	NI	nor [neither, not]
13	dans la connaissance	in the knowledge	V ZNANII	in the ZNANII *(knowledge)*
14	des calculatrices	of the electronic	ELEKTRONNYH	of electronic [electron]
15	électroniques,	calculatrices, *(computers)*	VYČISLITELNYH MAŠIN,	numeral MAŠIN *(digital computers),*
16	les problèmes	the present	AKTUAL'NYE	actual
17	actuels	problems	PROBLEMY	problem-s
18	de la traduction	of the automatic	AVTOMATIČESKOGO	of automatic
19	automatique	traduction *(translation)*	PEREVODA	translating
20	des	of the		of
21	langues,	langues, *(languages)*	JAZYKOV,	JAZYKOV *(languages)*
22	tel	such	TAKOVA	is such
23	est	is	EST'	is [there is, there are]
24	le but	the aim	CEL'	the purpose
25	dans	in	V	in [into, to]
26	lequel	which	KOTOROJ	which
27	ce	this	ETA	this
28	livre	book	KNIGA	book
29	a été	has been	BYLA	was
30	conçu.	conceived.	ZADUMANA.	ZADUMANA *(conceived).*
31	Nombreuses	Numerous	MNOGOČISLENNY	Are numerous
32	sont	are	SUT'	are
33	les difficultés	the difficulties	TRUDNOSTI	difficulty-s

	I *Author's Original French*	II *Brown's Translation, I.B.M. 704*	III *Corbe's Translation, manual*	IV *Swanson's Translation, I.B.M. 704*
34	qui	which	KOTORYE	which
35	se dressent	set up them-selves	VSTAJUT	VSTAJUT *(arise)*
36	encore	000017 *(still)*	EŠČE	yet [still]
37	sur	on	NA	on
38	le chemin	the path	PUTI	the way [means]
39	avant	before	PREŽDE	before
40	qu'		CEM	
41	une	a	ODIN	one [alone]
42	traduction	traduction *(translation)*	PEREVOD	PEREVOD *(translation)*
43	satisfaisante	satisfying	UNDOVLETVOR-ITEL'NYJ	satisfactory
44	pour	for	DLJA	for
45	un	a	ODNOGO	one [alone]
46	lecteur	reader	NEMNOGO	of not much
47	un peu	a little	TREBOVATEL'NOGO	TREBOVATEL'NOGO *(demanding)*
48	exigeant	requiring	ČITATELJA	reader
49	puisse	can	MOŽET	can
50	sortir	exit	VYITI	VYITI *(emerge)*
51	d'	from	iz	from [of]
52	une	a	ODNOJ	one [alone]
53	machine;	machine;	MAŠINY.	MAŠINY . *(machine)*
54	De grands	Large	BOLŠIE	Large
55	progrès	advances	PROGRESSY	PROGRESSY *(progress)*
56	ont été	have been	BYLI	were
57	réalisés	realized	DOSTIGNUTY	attained
58	dans	in	V	in
59	l'analyse	the analysis	ANALIZE	the analysis
60	des langues,	of the langues, *(languages)*	JAZYKOV,	of JAZYKOV *(languages)*
61	et	and	I	and
62	les principales	the principal	OSNOVNYE	*(the)* principal
63	étapes	steps	ETAPY	stages
64	de l'étude	of the study	IZUČENIJA	of the study
65	du langage	of the language	JAZYKA	of the JAZYKA *(language)*
66	en	in	s	with
67	vue	view	CEL'JU	the purpose
68	de la traduction	of the automatic	AVTOMATIČES-KOGO	of automatic
69	automatique	traduction *(translation)*	PEREVODA	translating

	I *Author's Original French*	II *Brown's Translation, I.B.M. 704*	III *Corbe's Translation, manual*	IV *Swanson's Translation, I.B.M. 704*
70	sont	are	SUT'	are
71	maintenant	now	TEPER'	now
72	définies.	definite.	OPRBDELENY.	are determined.
73	La recherche	The research	ISSLEDOVANIE	The investigation [research]
74	a progressé	has progressed	PROGRESSIROVALO	PROGRESSIROVALO *(has progressed)*
75	de façon	in a spectacular	EFFEKTNYM	by the EFFEKTNYM *(effectively*
76	spectaculaire	manner	OBRAZOM	way
77	depuis	since	S	from [with]
78	1955.	1955.	1955.	1955.
79	Des traductions	Useful	POLEZNYE	Useful [effective]
80	utiles	traductions *(translations)*	PEREVODY	PEREVODY *(translations)*
81	sont	are	SUT'	are
82	faites	done	DELAEMY	are maked [doing] *(made)*
83	par des machines	by machines	MAŠINAMI	MAŠINAMI *(by machines)*
84	et	and	I	and
85	leur nombre	their number	IH ČISLO	their number
86	ira	will	POIDET	POIDET *(will go)*
87	en croissant,	increase continuously,	VOZRASTAJA,	increasing,
88	leur	their	IH	their
89	qualité	quality	KAČESTVO	quality
90	s'améliorera	will improve itself	BUDET	will
91			ULUČŠAT'SJA	be improved
92	constamment.	constantly.	POSTOJANNO.	constant-ly.
93	Mais	But	NO	But
94	certaines	certaines *(certain)*	NEKOTORYE	certain [some]
95	découvertes	discovered *(discoveries)*	OTKRYTIJA	OTKRYTIJA *(discovery-s)*
96	sont	are	SUT'	are
97	nécessaires	necessary	NEOBHODIMY	are necessary
98	pour que	for	ČTOBY	that
99	cette	this	ETO	this
100	recherche	research	ISSLEDOVANIE	investigation [research]
101	entre	to enter	VOŠLO	VOŠLO *(enter)*
102	bientôt	soon		
103	dans		v	into [in, to]

	I *Author's Original French*	II *Brown's Translation, I.B.M. 704*	III *Corbe's Translation, manual*	IV *Swanson's Translation, I.B.M. 704*
104	une	a	ODNU	one [alone]
105	nouvelle	new	NOVUJU	new
106	phase,	phase,	FAZU,	phase,
107	celle	that	ETU	this
108	de l'automatisation	of the automatisation	AVTOMATIZACII	AVTOMATIZACII *(automation)*
109	à	at	NA	on
110	98	98	98	98
111	ou	or	ILI	or
112	99	99	99	99
113	pour cent.	pct.	PROC.	PROC. *(per cent)*
114	Le lecteur	The reader	ČITATEL'	The reader
115	qui	which	KOTORYJ	which
116	souhaite	*souhaite (wishes)*	ŽELAET	desires
117	se tenir	hold	DERŽAT'SJA	DERŽAT'SJA *(to keep abreast)*
118	au	at	V	in [into, to] *(to keep abreast)*
119	courant	the current	KURSE	KURSE *(to keep abreast)*
120	et	and	I	and [also]
121	suivre	follow	SLEDOVAT'ZA	to follow during
122	cette	this	ETIM	these [this]
123	future	futide *(future)*	BUDUŠČIM	future [willing, will be]
124	étape	step	ETAPOM	stage
125	du progrès	of the advance	NAUČNOGO	of the scientific
126	scientifique	scientific	PROGRESSA	PROGRESSA *(progress),*
127	trouvera	find	NAIDET	will find
128	ici	here	ZDES'	here
129	une	a	ODNO	one [alone]
130	mise au point	put at the point	UTOČNENIE	refinement
131	de l'état	of the state	AKTUAL'NOGO	of the actual
132	actuel	present	SOSTOJANIJA	state [condition, position]
133	des travaux	of the work	RABOT	of works [papers],
134	tel qu'	such that	TAKOGO	such
135	il	it	KOTOROE	which
136	ressort	be evident	VYTEKAET	flows out [follows]
137	des	from the	IZ	from [of]
138	ouvrages	work	TRUDOV	works [treatise, difficulty]
139	et	and	I	and
140	articles	article	STATEJ	articles

	I *Author's Original French*	II *Brown's Translation, I.B.M. 704*	III *Corbe's Translation, manual*	IV *Swanson's Translation, I.B.M. 704*
141	parus	paru *(published)*	POJAVIVŠIHSJA	of appearing
142	depuis	since	S	from [with]
143	1955	1955	1955	1955
144	aux	at the	V	in [into, to]
145	Etats	state	SOEDINENNYH	connected *(United States)*
146	Unis,	unite, *(United States)*	ŠTATAH	ŠTATAH *(United States)*
147	en	in	V	in [into, to]
148	Grande-Bretagne	large-britain *(Great Britain)*	VELIKOBRITANII	VELIKOBRITANII *(Great Britain)*
149	et	and	I	and
150	dans	in	V	in
151	l'Union	the union	SOVETSKOM	by SOVETSK-
152	soviétique	sovietic. *(Soviet Union)*	SOJUZE.	SOJUZ. *(Soviet Union)*

Index