Progress in automatic language translation

by **ANDREW D. BOOTH** D.Sc.
University of London

THE FIRST INTERNATIONAL CONFERENCE ON THE SUBJECT of machine translation was held at the Massachusetts Institute of Technology in 1952. All of the known experts were present, together with sundry observers, but the whole conference was comfortably housed in a small committee room. The discussions which took place at that meeting are now historical. They resulted in the volume 'Machine Translation of Languages', edited by A. D. Booth and W. N. Locke, and some of the opinions expressed therein have no doubt caused embarrassment to the original speakers many times since the conference ended. A second M.I.T. conference was held two years later, and this was followed in 1959 by a conference held at Western Reserve University, Cleveland, on the subject of standards for a common language in machine translation. The latter conference, too, has been well documented in two volumes which have just appeared. Not unnaturally, so ambitious a scheme as the standardization of techniques within a growing field was doomed to failure. The arguments which took place at this conference, however, are of some permanent value, and, when the time is fitting to consider such standards, may have considerable influence on the discussions.

Last month there was held, at the National Physical Laboratory, the first of these international conferences on machine translation to take place in this country. Over thirty-five papers were read, some by well known pioneers, others by the newcomers who are making such progress at the present time.

HOW IT ALL BEGAN

Before making any attempt to discuss any of the papers in detail, it is worth considering the history of machine translation and the area in which activity appears to be greatest at the present time.

Apart from the early patent proposals of Trojanskij, which were not apparently based on any physical mechanism, the first practical proposals for translation by machine were made by the present author in 1946. At that time they amounted, effectively, only to the use of a computing machine store as an automatic dictionary. Even in this respect the early proposals lacked real practicality. The dictionary proposals would have required storage capacities greater than those even now available, and certainly greater than anything which was envisaged in 1946. These difficulties were not unknown to the research workers, and, in fact in 1947, the first practical proposals in translation were put forward by Booth and Richens, and were tested, both by simulated experiments using human operators performing in the sort of blind way that a computer operates, and by experiments on standard punched-card machinery. In essence, a dictionary was used in these experiments for storing, not whole words, but the decomposition of these words into stems and associated endings. The now well known argument, that from N stems and M endings, $M \times N$ words can be formed, for the storage of only $M + N$ linguistic items, shows how improved thinking could lead to possible means of using the crude and imperfect computers of the early 1950s. The first Booth-Richens proposals intended either to ignore the effect of inflexion altogether, or alternatively to have as the machine output the basic meaning or meanings of words, and some grammatical notes to assist the reader in making sense of the jumble. This tacit assumption that the reader should make sense of the machine translation was in fact the same as the suggestion which Reifler made slightly later, that machine-translated text would need 'post-editing'. Reifler, however, improved on this scheme by suggesting not only a post-editor, but also a pre-editor, who should be an expert in the language from which translation is to be made, and who should remove ambiguities from the incident text. All of these ideas were well thrashed out at the 1952 conference, and when, at Birkbeck College in 1955, the work was taken up in earnest, under the auspices of the Nuffield Foundation Grant, one of the first things to be done was to modify the original Booth-Richens scheme in such a way that stems and endings in the incident language were transformed into stems and endings in the language translation, so that sensible text was produced on a word-for-word basis, even though, considered in a wider context, much might be desired. At the same time, that is, in 1955–56, the first trials of a new machine-translation program for French to English were made on the Birkbeck computer. This program, quite apart from the fact that it re-united stems and endings in translation, also processed material sentence by sentence, rather than word by word; and furthermore as has been clearly shown by the present author,* the method implied what is now known as 'predictive analysis'. Of this it is sufficient to say that, in operating it, one makes certain assumptions at the very start of a sentence about the way in which the structure of that sentence will proceed. These assumptions are being continually checked as new words become available, so that finally a correct, or as nearly as possible correct, version is produced. In this country at least, progress from this point onwards has been very slow. The reasons for this are twofold. First, and most important, the linguistic data required for writing comprehensive machine programs, even for French to English translation, are almost completely lacking. Second, the machines available in this country have, up to the present, been completely inadequate for any large-scale experiments in machine translation.

To deal with the first point, we may say at once that, upon realizing the defects of classical linguistic knowledge, the efforts of the team at Birkbeck College were immediately directed to methods for the automatic analysis of text, and these investigations have led not only to automatic means of making dictionaries, but also to means of analysing text for stylistic differences. This work was pioneered by L. Brandwood in the mid-1950s, and has since acquired great notoriety, the latest instance being the use of his methods to establish the homogeneity of the existing works of Homer.

In the United States there were some ill-advised experiments, designed more for publicity than for real scientific progress, after which a number of groups settled down to solid work. The pioneers in automatic ditionary-making in the United States were the Harvard Computational Laboratory, where Anthony Oettinger and his team have done sterling work on Russian, which has recently been reported in a large volume. At the Massachusetts Institute of Technology as well, a group under Yngve has produced a number of ideas about the basic structure of all languages. Probably the most promising of these is the idea of 'depth' of language, which may have important repercussions on the futur Starting somewhat later, the National Bureau of Sta dards, after investigating some of the Georgetov University programs, broke new ground for themselv when Ida Rhodes rediscovered the idea of predicti analysis, which has been briefly explained above. C this side of the Atlantic the idea of predictive analysi perhaps appears to be so self-evidently necessary tha it requires no justification. But looking at the con temporary literature in the United States, it certainl appears that, at first at least, there was some difficult in obtaining acceptance of the virtues of this technique Other groups of workers in the United States have als been active in the field, and they will be mentioned b name in the detailed report which follows. But suffice it to say that, although sound work is being done b these groups, it does appear at the moment that materia of such originality has appeared as to form new pointer to the future.

Russian workers, too, have been most active, the groups in Moscow under Ivanov and in Leningrad under Andreev, to say nothing of the active nucleus of workers in Kiev, have added greatly to the theory of Meta-languages, and it is a great pity that the Russians did not submit papers to the conference. Most of the available reports of Russian work are either out of date or are partial accounts gleaned from second-hand conference reports.

It is perhaps unpopular to urge international co-operation at the present time, but it would certainly be helpful if the Soviets would see that their workers were enabled to attend international conferences, and that such attendance was notified well in advance. It is a source of embarrassment to the organizers of conferences when the Russian delegates arrive unexpectedly two days after the conference has started.

WHAT WAS REPORTED AT THE N.P.L. CONFERENCE

We now come to last month's National Physical Laboratory conference. Thirty-five papers were circulated. These were selected from many more submitted to the International Committee which acted as referee. If it were necessary to say, in a single sentence, the way in which the conference pointed to the future of machine translation, it would be this; that the important fields of activity at the present time are in automatic syntactic analysis and in predictive analysis as a means of accurate translation.

Linguistic relationships

At the conference Paul Garvin (Ramo-Woolridge Corp.) read a paper on the heuristic aspects of automatic linguistic analysis. He examined the relations of language analysis, to such things as the analysis of games. In the latter, with a self-organizing machine, a simple target can be prescribed, namely that the object is to win. This does not occur, however, in language translation, where

* Journ. I.E.E., 3 (1957) 629

the far more nebulous objective is to produce the best possible translation. Garvin's experimental work is as yet in an early stage and does not appear to have been tested on any computing machine. In effect, his earliest experiments will merely compare input text with the dictionary, examine the environments of words in that text, and detect the occurrence of frequent juxtapositions of different letter or word combinations. This can of course, be quite a useful adjunct to more powerful analyses, but it remains to be shown whether this approach will be a fruitful one.

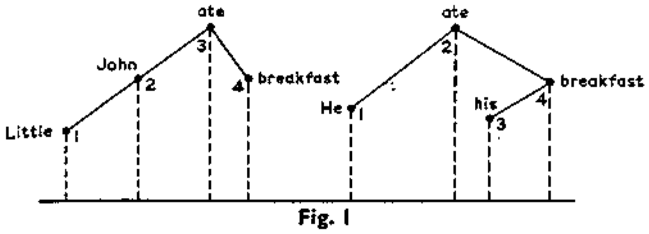

Fig. 1

Warren Plath (Harvard) considered the problem of automatic sentence diagramming. This may appear a novel idea to the uninitiated, but, effectively, the suggestion is that the various words in a sentence can be put into a relation with various forms of tree-like structure, as shown in Fig. 1. The important thing about such tree structures is that they enable one to classify languages, and in particular to decide whether languages are ' projective ' or ' non-projective '. It appears, from the work which has been done up to the present, that English, French, German and Russian are projective languages, and this, if it can be established rigorously, is a matter of great importance for machine translators.

These graphical ideas of language structure really stem from the early pioneering work of the Modern Language Research Unit at the University of Cambridge, who were represented by an important paper by A. F. Parker-Rhodes, one of the founders of the group. He suggested a new model of syntactic description, which leads to a satisfactory method of bracketing words in a sentence so as to indicate their functional dependence upon each other. An example of this is the following:

((A (rather lazy)cat)(chases(falling(leaves and butterflies;)))) of course, they(can(easily get away.)))

It is interesting to notice that this bracketing procedure is, in fact, an alternative way of looking at the type of situation described by Plath. A second contribution from the Cambridge group was that by Sparck-Jones, who considered the problems of mechanized semantic classification, and attempted to formulate rules by which text can be automatically used to obtain improved data about classificatory methods.

Dealing with Russian

David Hays (Rand Corp.) spoke on the value of dependency connexions. His paper attempted to show how numerical values could be assigned to syntactic relations in such a way that relations of higher order are established in preference to those having lower values. By higher, in this context, is meant the same sort of thing as is indicated in the diagrams which have just been given. Also from the Rand Corporation was a paper by K. E. Harper on procedures for the determination of distributional classes. It transpired from this paper that the Rand Corporation workers have a corpus of 250,000 Russian words drawn from physics texts upon which they are able to make experiments. The particular experiments described by Harper aim at reducing the structures found in language syntax to terms similar to those which Plath had been discussing. A third paper from the Rand Corporation was that by Dean Worth, who gave transformation criteria for the classification of predicative genitive constructions in Russian. This paper which in a sense expresses everyday linguistics in mathematical logical terms, gives practical transformation rules for the Russian language, when it is considered for automatic translation into English.

Random sentence generating

The contributions from the Massachusetts Institute of Technology came from several departments. Victor Yngve described, in most interesting terms, his experiments on the random generation of English sentences. It may be enquired what use such work is in furthering the aims of machine translation, but a little reflexion will show that, if rules can be formulated which give grammatically correct sentences, these same rules are likely to form at least a part of the corpus required to define the structure of the languages themselves. Yngve formulated a set of 77 rules, which he derived by considering ten sentences in English, the first two of which were:

' Engineer Small has a little train.' ' The engine is black and shiny.'

During the past years, Yngve, realizing the difficulty of transcribing linguistic programs onto machines designed primarily for doing arithmetic, has been responsible for producing a linguistic auto-code called Comit. Using this auto-code and his set of 77 rules, his machine produced large numbers of sentences, which are given as examples in the paper. Some of the more amusing of these are:

' He has four polished sand-domes.' ' Water is big.' ' It's steam is proud of wheels.' ' He is oiled.'

Now all of these sentences are either meaningless or comical, but proper inspection shows that they have the merit of being correctly constructed sentences. It shows also, what is often not realized by the proponents of information theory, that information theory has no connexion with meaning in language, as indeed statistical treatments must inevitably fail to have. This does not imply, however, that the sentences produced by Yngve are useless, because meaning was not the object of the exercise. It would have been quite possible to produce meaningful and not particularly humorous sentences by adding a small sub-unit to Yngve's original program.

To be continued

Progress in automatic language translation

by **ANDREW D. BOOTH** D.Sc., *University of London*

Logic and structure

The next of the papers from M.I.T. was that by Elinor Charney on 'The Semantic Interpretation of Linguistic Entities that Function Structurally'. This somewhat ponderous title conceals a paper which is effectively about symbolic logic and its application to the analysis of linguistic terms of the conventional sort, such as *either, neither,* and so on. It may be argued that such abstract investigations will not soon lead to practical procedures in translation, but this would be a very false assumption, since mathematical logic is in fact a self-consistent language, to which it is hoped, in the future, natural languages may be reduced.

Edward Klima, also from M.I.T., spoke on structure at the lexical level, and its implications for transfer grammar. His paper was, in effect, a discussion of the differences between such words as *learn* and *know,* in phrases like 'learn a word' and 'know a word', and contained a conscientious attempt to classify such words for machine translation. A most valuable activity, since it is by such classifications that ambiguities and other difficulties will eventually be resolved.

The final contribution from M.I.T. was that by G. H. Matthews, on analysis of synthesis of sentences of natural languages, which was, in fact, a rival version of the predictive analysis technique.

The sister institution to M.I.T. in Boston is Harvard. This University, too, was well represented at the Conference. G. Salton and R. W. Thorpe described an approach to the segmentation problem in speech analysis and m.t. It is not generally realized that one of the great difficulties in the automatic analysis of peech by machine is the apparently trivial problem of deciding where one word ends and the next begins. The same sort of thing applies in the analysis of written text by machine, and Salton and Thorpe considered the use of grammatical indicators and a form of the ubiquitous predictive analysis.

Russian again

William Foust and Julia Walkling, also from Harvard, described a preliminary structural transfer system which was, in effect, a set of transformations which map Russian constructions onto English constructions. Since the discussion was only preliminary, the constructions were described only in very general terms. But sufficient was said to show that the method might with development prove of great utility.

Irina Lynch read a distinctly earthy paper on the Russian ся verbs, impersonally used verbs, and subject-object ambiguities. She provided something which was rare among the papers at the Conference, namely a flow chart, which showed in detail the decisions which must be taken in order to attain an accurate analysis of the Russian text.

Four stages for Japanese

The last of the Harvard contributions, and in many ways the most interesting, was that of Susumu Kuno, who described a preliminary approach to Japanese-English automatic translation. He suggested that four stages were necessary: automatic input editing, automatic segmentation with morphological analysis, syntactical analysis, and transformation with output editing, including semantic transfer.

We have here space to deal only with the first problem—the problem of pre-editing the input text to a form suitable for machine. Japanese characters are of two types: Kanas, which are the equivalent of Japanese letters, and are 71 in number; and Kanjis, which are effectively ideographs and of which there are eighteen hundred and fifty. Kuno's procedure is to code the Kanas into two-letter Roman alphabetical groups and to code the Kanjis using a straightforward binary or I.B.M.-type code. After this preliminary coding, the next problem is one of segmentation, just as it is in the automatic analysis of spoken dialogue. The problem arises because Japanese writing contains no spacing. The method which Kuno proposes is to match dictionary stems and find the longest of these which start the lines of the Japanese text. These stems having been deleted, possible endings are hunted, and processed, and in this way the complete character is isolated and removed from the text. This having been done, the next character is considered, and so on.

In effect Kuno's method is an extension of that proposed originally by Booth and Richens for the analysis of German compounds, and presents few new features. It is, however, most interesting to see that these techniques can be applied to a language like Japanese, which is apparently far removed in structure from that of the European languages.

The Japanese themselves are not inactive in the field of machine translation, and one paper was presented by a Japanese worker, I. Sakai. This dealt with syntax in universal translation, and described proposed experiments in translation, using the Japanese parametron computer, which has among its components cores, drums, and three magnetic-tape units. It is too early to say what results may be expected from these experiments, but it will be interesting to see the progress which the Japanese make.

Mechanized syntax

The Berkeley branch of the University of California is well known for the productions of Sydney Lamb. On the occasion of this conference he read a paper on the mechanization of syntactic analysis. This described a simple program which listed the different words in a 5000-word text, selected from the writings of Winston Churchill, counted the occurrences of words, and then worked out so called ' left neighbour ' and ' right neighbour ' counts. That is, the number of times that the given word in the text occurs, divided by the number of times that some other word occurs to the left or right of it. A number of examples of such counts are given and some indications are produced which show, firstly the essential difficulty of doing language statistics on any words apart from nearest neighbours, and secondly of the way in which even this limited information can be of great help for machine translation.

Chemical terminology

J. H. Wahlgren, also from Berkeley, discussed the linguistic analysis of Russian chemical terminology. The paper was an interesting and earthy one, which attempted to describe the way in which Russian chemical phraseology works. The description, quite apart from its use in machine translation, should be of great value to anyone who is working in the translation of Russian chemistry.

Dostert's group, from Georgetown University, submitted several papers, and, among these, one by Lawrence Summers discussed the machine translation of Russian organic chemical names. The subject matter was similar to that of Wahlgren's paper just mentioned, but, in the case of Summer's analysis, the argument proceeded by an analysis and re-synthesis of the component fragments of Russian names. He pointed out that the number of possible organic compounds is effectively infinite, so English equivalents of Russian words cannot be stored as a whole. The paper gives a table of a hundred fragmentary equivalents, and a set of rules for converting Russian to English from these. The basic technique is again the old stem-ending decomposition one suggested by Booth and Richens for German compounds.

Obscurity and clarity

The other contribution from Georgetown was that by Michael Zarechnak, who suggested that a fourth level of linguistic analysis was needed to improve the Georgetown programs for Russian–English. At the present time these have three levels of analysis—morphological, syntagmatic and syntactic. The new analysis, which, frankly, I found not entirely comprehensible, was apparently shown to be necessary by the results of practical experiments. Anything indicated by experiment must necessarily be taken seriously, and it is to be hoped that when full publications of Zarechnak's ideas become available, a useful contribution will result.

The Bureau of Standards has already been mentioned as one of the active institutions in the field of machine translation. Franz Alt and Ida Rhodes jointly contributed a paper on the recognition of clauses and phrases in the machine translation of language. Unlike some of the papers this one was very well written and understandable, a pleasure to read. It suggested a simple algorithm which enables the machine to decide when clauses start and when they finish, and also to see if any clauses are left incomplete at the end of a sentence, in the latter event, of course, the machine returns to the beginning, and repeats its analysis *via* a different route until a correctly terminated sentence is obtained. The paper, apart from its valuable idea content, contained a detailed analysis devoted to Russian and gave numerous examples both in Russian and in English.

The United States Air Force, well known for its support of machine translation, and for its construction of large-scale dictionary mechanisms, produced a paper by Murray E. Sherry on the identification of nested structures in predictive syntactic analysis. The idea of nested structures was illustrated earlier in the work of Parker-Rhodes. Sherry's contribution showed how, by the use of certain sentinel words, clause determination could be assisted and predictive analysis rendered more simple.

L. R. Mickelsen (I.B.M. Corp., U.S.A.) considered source language specification with table look-up in a high-capacity dictionary. He discussed the well known idea of a dictionary containing words with associated structure numbers originally defined in the work of the Birkbeck College group. His contribution was to apply this idea, explained in detail originally for French-English, to the combination Russian-to-English.

Continental contributions

Of the European workers, the University of Milan provided a strong contingent. Silvio Ceccato and Runa Zonta, of the University of Milan, attempted an analysis of how humans go about translation in a paper entitled ' Human Translation and Translation by Machine '. Such an analysis of human mental processes is quite likely to form a basis for any large-scale work on machine translation, and the work of Ceccato and Zonta will be read with interest by all workers in this field.

From Paris, P. Meile considered problems of address in an automatic dictionary of French. His paper described an attempt to economize dictionary space by giving the first N letters of a word (he suggested the value $N=6$) plus the total number of letters, or a terminal letter or the ending itself. Such ideas for word length compression are, of course, quite old. They were used both at Birkbeck College and in the Russian work. On the other hand, the particular scheme suggested by Meile has certain attractions, and merits consideration.

M. Corbe and R. Tabory (I.B.M., France) described an introduction to an automatic English syntax by fragmentation. This contained an interesting analysis of the proportion of unambiguously recognizable parts of speech of the words in a text. The paper provided numerous statistics, drawn from a sample containing about a thousand words, and the fact that, on average, something like one third of the words are grammatically ambiguous will come as a surprise to many people who have not been familiar with this problem in practical experiments on machine translation.

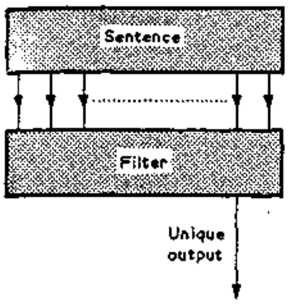

Fig. 2

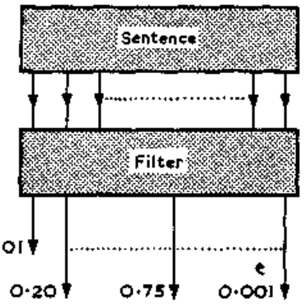

Fig. 3

Probably the most interesting of all of the European contributions was that by Y. Lecerf, from Euratom. His paper, entitled 'Intrinsic Machine Addressing in Automatic Translation', contained an analysis of the difficulties of using structural information. He considered the example:

Le	**page**	**brise**	**la**	**pointe**	**de**	**la**	**lance**
Art.	*N. fem.*	*V*	*Art.*	*V*	*Prep.*	*Art.*	*V*
Pr.	*N. masc.*	*N*	*Pr.*	*N*		*Pr.*	*N*
			N			*N*	

This example has, considering the possibilities of arranging the different parts of speech in the order of words, 2×2×2×3×2×1×3×2 = 288 possibilities: He suggested, first of all, that the correct member from this set of possibilities might be found by an automaton having the structure shown in Fig. 2. This is unlikely to be practically useful because of the complexity of the filter, which has very many inputs and a single output. Furthermore, it is unlikely that for even simple sentences the filter would give an unique output. What is really needed, Lecerf pointed out, is a filter which gives the possible outputs and accompanies these by the probabilities that they are the true output. This idea is shown in the filter diagram reproduced in Fig. 3. Unfortunately, for practical reasons, this scheme, too, would be unworkable because of difficulties in constructing the filter, so that a set of sub-filters, each dealing with a part of the sentence which can be treated by itself, was substituted, which fed a full filter giving the correct probabilities. This discussion of filters is interesting in itself, but Lecerf's long paper went on to discuss certain topological analogies between language and the geometry of curves, and suggested that the idea of the intrinsic equation of a mathematical curve might provide a clue to an invariant description of language.

British work

We come now to the British contributions. There was a paper from the National Physical Laboratory by D. W. Davies and A. M. Day on a technique for consistent splitting of Russian words. This contained a detailed analysis of the stem-ending decomposition process, described in terms suitable for application with the National Physical Laboratory's version of the Harvard dictionary and the N.P.L. computers Ace to Deuce.

A second paper from N.P.L. was that by J. McDaniel and S. Whelan on the grammatical interpretation of Russian inflected forms using a stem dictionary. The title of this paper is self-explanatory but it is worth remarking that the paper formed a distinct contribution to the detailed linguistic literature of this subject, and will undoubtedly be of considerable use to all those who deal with Russian to English translation by machine.

Birkbeck College was represented by a single paper read by Michael Levison, on the mechanical analysis of language. This described some actual machine applications to glossary construction and letter-group frequency analysis. It then proceeded to analyse the efficiency of different methods of list construction by computer, an important topic since, even at the present time, computer storage is relatively limited, and the time wasted if lists of words have to be displaced for the insertion of new data can be considerable.

Summing up

The conference discussions were stimulating and often heated. Nevertheless, this conference can be counted a distinct success. The papers presented, and also an edited version of the discussion, will be published in due course by the Stationery Office. All workers in the field who did not have the pleasure of attending the conference will certainly wait with considerable anticipation for the production of this volume.

The organizers of the conference are to be congratulated upon this smooth-running piece of work. Perhaps the fact that the Chairman was Albert Uttley, Superintendent of the Autonetics Division of N.P.L., and well known for his work on self-optimizing automata, may be a reason for this good organization, for surely, if Uttley can construct a self-optimizing automaton, he must indeed be self-optimizing himself. And, this being the second of such conferences that he has organized, shows that he lives by his own precepts.