[*Endeavour*, vol.27, 1968]

Machine translation— The second phase of development

Sergei Perschke

The practical realization of the relatively new art of machine translation has been made possible only through close collaboration between linguists, mathematicians, and computer technologists. A measure of the success achieved is the fact that the Foreign Technology Division of the United States Air Force is already translating mechanically some 100 000 words of Russian technical text into English per day. The author reviews the historical development of machine translation against the background of the work that has been carried out by Euratom at Ispra in Italy. Here the stage is now being set for more sophisticated forms of machine translation based on recent advances in computer design and more profound studies of the linguistic and semantic problems involved.

The first phase

There are two main reasons for the evolution of machine translation. The first is what is now familiarly described as the information explosion, and the second is the rapid advance in computer technology which has led to the mechanization of more and more human activities.

Nearly a quarter of the abstracts published in abstracting journals such as *Nuclear Science Abstracts* and *Chemical Abstracts* are of articles drawn from Soviet scientific periodicals. The growing importance of scientific work in the eastern European countries makes ever more pressing the need for translation of Russian documents into English. Shortage of qualified translators and rising labour costs make it more and more difficult to meet the increasing information needs of the scientific communities of the western world.

Translation was first examined as being potentially susceptible to mechanization in the late forties and the . early fifties with very promising results. Following a series of theoretical discussions, the first experimental translation programme was carried out as a joint venture by International Business Machines and Georgetown University, Washington. The programme was completed at the end of 1953 [1]. It consisted of a few hundred words and enabled a small selected text to be translated. The excessive publicity given to this experiment created the belief that machine translation was not very far from becoming a reality.

The problems of machine translation are essentially linguistic. To produce a translation it is not enough, as was announced after the 1953 experiment, merely to compile a large machine-readable bilingual dictionary and to formulate a few linguistic rules that identify the differences between the 'source' and 'target' languages. It was soon realized, however, that existing knowledge of the basis of language was still too scanty to permit a more sophisticated method of automation.

The approaches to the solution of the problem of machine translation at this point had to diverge. Those whose aim was to produce practically usable automatic translations in as short a period as possible had to forego basic research and their methods remained empirical, and were later dubbed the 'brute-force approach'. Theirs may be called the dictionary approach. Translation was defined as a substitution of signs. Because pure word-for-word translation, that is, the substitution of one English word for one Russian word, turned out to be unsatisfactory, it was refined by introducing into the dictionary not only the single words, but also phrases.

It is now obvious that this relatively primitive method was bound eventually to become incapable of further improvement, because the number of dictionary entries grew too rapidly for any improvement in translation quality to be possible, and the linguistic knowledge was not sufficient to replace this method efficiently by a syntactic and semantic analysis of language. Nevertheless, several projects succeeded in developing operational systems of machine translation, of which two, realized by IBM and Georgetown University, arc being used for practical translation services on a fairly wide scale.

The system developed by IBM is an example of a pure 'brute-force approach'. It consists of a special-purpose computer constructed around a particular storage device—the photoscopic disc—which permits a reasonably fast direct access to the dictionary entries. The system is called the 'bi-directional single-pass translator', and theoretically could translate both from Russian into English, and from English into Russian. The IBM system became operational in 1963. It translates approximately 10 000 words per hour and is at present being used by the FTD (Foreign Technology Division of the U.S. Air Force) on a very large scale (about 100 000 words per day). However, in order to improve the quality of translation, the machine output is normally given some subsequent human editing.

The system developed by Georgetown University is an example of a modified 'brute-force approach'. The technical characteristics of the second generation of computers did not permit the construction of a large dictionary with the direct access capabilities of the photoscopic disc which had never been available with general-purpose computers. The only storage medium for the dictionary was magnetic tape, which imposed a sequential access upon the dictionary entries. The words occurring in the text to be translated had therefore to be sorted alphabetically before dictionary search and put back into the original word order afterwards. In order to obtain results equivalent to or better than those with the IBM system, it was necessary to introduce at least rudimentary algorithms of syntactic and semantic analysis. By the end of 1961 the system had become operational, and it is the first and only one that operates with a general-purpose computer for a practical machine translation service.

Sergei Perschke
Was born in Sablino (Leningrad) in 1936, and studied at the University of Cologne. He joined the *Centro di Cibernetica* of the University of Milan in 1959 to participate in the machine translation project as a linguist, and since 1965 he has been responsible for the Russian-English machine translation service and development at Euratom's Scientific Data Processing Centre (CETIS) at Ispra in North Italy.

Both of these systems were criticized on the grounds of shortcomings in the theoretical treatment of the linguistic problems. However, viewed in retrospect, it seems certain that they were the only way of achieving practical results in a reasonably short time. Those projects which had attempted more ambitious solutions failed, as far as the practical results are concerned, because of the multiplicity of basic problems and the incompatibility of traditional grammars with mechanization.

Most attempts at a theoretically satisfactory solution of the problem of linguistic analysis aimed at a description of language by mathematical methods. N. Chomsky's work on mathematical linguistics [2] was taken as a basis for attempts to obtain an exact description of linguistic structures. However, in these approaches, the basic error was frequently committed of confusing the mathematical model of a formal language with the syntactic description of a natural language. This confusion was further aggravated by the fact that mathematical logic uses linguistic terminology (for example the expressions 'word', 'proposition', 'syntax', 'semantics', and so on) and also, for illustration purposes, examples taken from natural language. Thus, attempts at using mathematical language models in linguistics digressed either into purely theoretical research such as the generation of random sentences, operating within the limits of mathematical logic without claiming a solution of linguistic problems, or they led to a 'pseudo-science' in which a complex mathematical apparatus was built up to describe the most elementary linguistic facts.

Another attempt at resolving the problem of mechanizing language was made at the University of Milan [3]. It started from an original approach to the philosophical problem of knowing and tried to describe human mental activities in terms of elementary, mechanizable operations. This approach did not produce practical results either, but a very good insight into some problems of linguistic relations was obtained, and the graphical representation of syntactic relations, used in the second part of this article is taken from this project.

Machine translation at EURATOM

Shortly after the establishment of the European Scientific Data Processing Centre (CETIS) by Euratom in 1959, machine translation was first considered within the wider context of documentation. But the American projects were already so far advanced as to make it appear impracticable to start development *de novo.* By 1961, when the installation of a large computer (IBM 7090) at CETIS at Ispra in North Italy made machine translation feasible from the technical point of view, the Georgetown system was already almost operational. One of its advantages consisted in the fact that it too was programmed for an IBM 7090 computer and thus could be run at Ispra without appreciable modification.

The first contacts with the Georgetown University project were established during the Symposium on Machine Translation at the National Physical Laboratory, Teddington, England, in 1961. In the following year, the system was put at the disposal of CETIS so that it could be used at Ispra to provide an experimental machine translation service and to test its practical utility. On the basis of this initial experience, which was promising despite some criticism of the linguistic background, a research contract was concluded with Georgetown University in 1963 putting the Russian-English machine translation system at the disposal of CETIS after making a number of modifications to suit the particular requirements of the Ispra establishment.

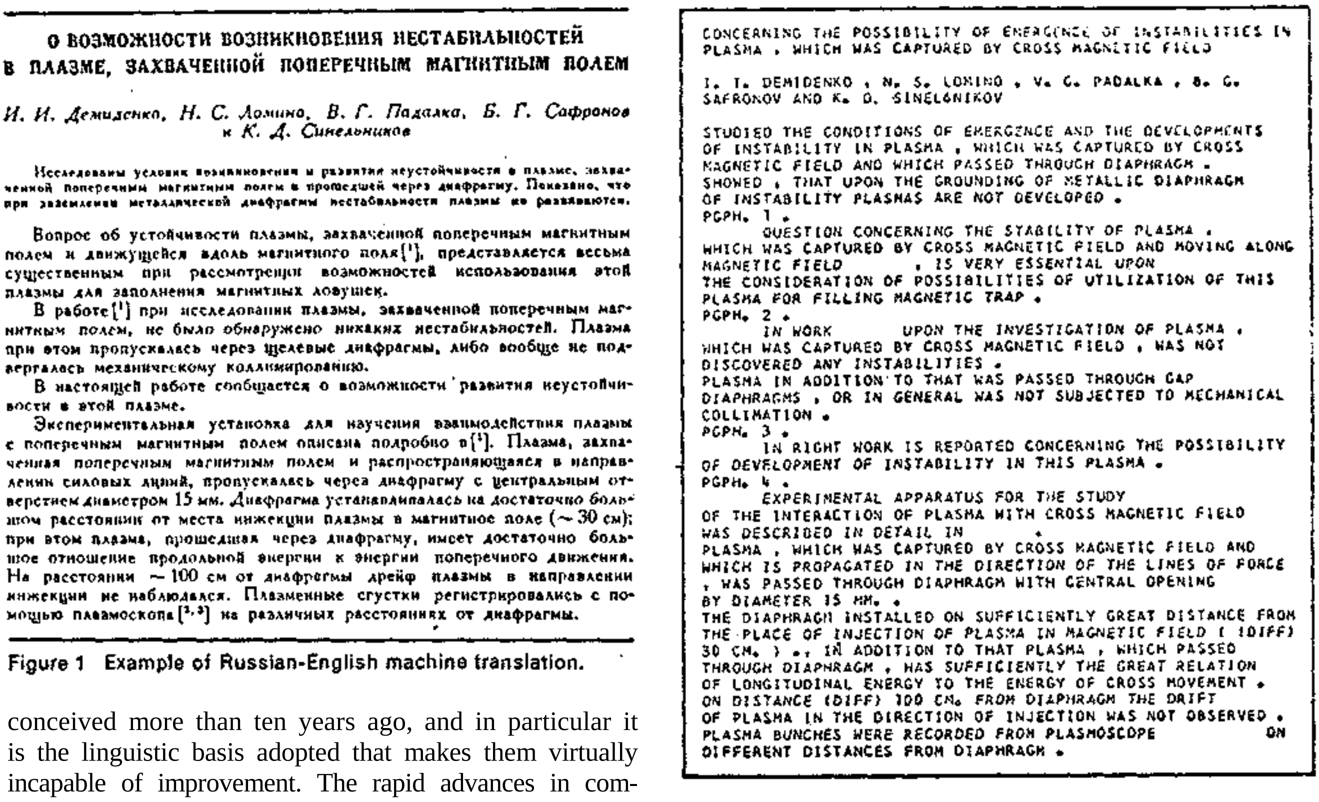
The primary purpose of the acquisition of the Georgetown system by CETIS was to provide the scientists at Ispra with a rapid and economic Russian-English translation service. Following the 1963-64 probationary period, when time was needed for organizing administrative and keypunching services, requests for translations of Russian texts have increased year by year. To provide a rapid 'awareness' service of incoming Russian publications, since 1966 contents lists of the periodicals have been translated automatically and distributed at the Centre. This initiative was well received, with the result that not only has the number of requests increased considerably (from about 30 in 1965 to some 120 in 1967), but the proportion of recent publications translated has risen from 12 to some 70.

In view of the success of this 'awareness' service, we are planning to translate the abstracts as well as the titles of incoming articles in the near future. Translation of abstracts is the first step towards a much wider application of machine translation in the field of documentation. An experiment in the automatic allocation of English keywords to Russian documents was performed early in 1967 with encouraging results [4]. The advantage of this application lies primarily in the promptness with which access is obtained to the original Russian documents within the framework of a large documentation and information retrieval system, since the recording of foreign articles in western abstract journals is normally subject to a delay of at least six months. Moreover, indexing becomes a by-product of automatic translation at virtually no extra cost. Further development of this experiment depends upon the results of another project in hand at CETIS, to study the optimal methods and strategies of automatic assignment of keywords.

At Ispra, the translations are delivered to the customers without post-editing. Although the quality of the translations is not perfect, it seems adequate for information requirements, since so far customers have not made use of the alternative facility offered of a 'human' translation of the text should they be dissatisfied with the automatic translation. It is difficult to assess the quality of the translations, since no exact methods of measuring quality exist at present, and the reading and comprehension tests used are not very reliable. However, when judging the usefulness of machine translation, speed and economic factors are of paramount importance. The present cost of machine translation of approximately 7 dollars per 1000 Russian words (excluding overhead costs and capital repayments on the development work, but including the capital repayments on the computer and keypunch equipment staff costs) would justify machine translation even if part of the texts had later to be retranslated by human means. Despite the adverse report [5] by the Automatic Language Processing Advisory Committee of the U.S. National Academy of Science in 1966, the success we have already achieved at Ispra encourages us to redouble our efforts to improve existing systems. Figure 1 illustrates the quality achieved.

The second phase

The machine translation systems at present in use were



О ВОЗМОЖНОСТИ ВОЗНИКНОВЕНИЯ НЕСТАБИЛЬНОСТЕЙ В ПЛАЗМЕ, ЗАХВАЧЕННОЙ ПОПЕРЕЧНЫМ МАГНИТНЫМ ПОЛЕМ

И. И. Демиденко, Н. С. Ломино, В. Г. Падалка, Б. Г. Сафронов и К. Д. Синельников

Исследованы условия возникновения и развития неустойчивости в плазме, захваченной поперечным магнитным полем и прошедшей через диафрагму. Показано, что при заземлении металлической диафрагмы нестабильности плазмы не развиваются.

Вопрос об устойчивости плазмы, захваченной поперечным магнитным полем и движущейся вдоль магнитного поля[1], представляется весьма существенным при рассмотрении возможностей использования этой плазмы для заполнения магнитных ловушек.

В работе[1] при исследовании плазмы, захваченной поперечным магнитным полем, не было обнаружено никаких нестабильностей. Плазма при этом пропускалась через щелевые диафрагмы, либо вообще не подвергалась механическому коллимированию.

В настоящей работе сообщается о возможности развития неустойчивости в этой плазме.

Экспериментальная установка для изучения взаимодействия плазмы с поперечным магнитным полем описана подробно в[4]. Плазма, захваченная поперечным магнитным полем и распространяющаяся в направлении силовых линий, пропускалась череа диафрагму с центральным отверстием диаметром 15 мм. Диафрагма устанавливалась на достаточно большом расстоянии от места инжекции плазмы в магнитное поле (~ 30 см); при этом плазма, прошедшая через диафрагму, имеет достаточно большое отношение продольной энергии к энергии поперечного движения. На расстоянии ~ 100 см от диафрагмы дрейф плазмы в направлении инжекции не наблюдался. Плазменные сгустки регистрировались с помощью плазмоскопа[2,3] на различных расстояниях от диафрагмы.

```
CONCERNING THE POSSIBILITY OF EMERGENCE OF INSTABILITIES IN
PLASMA , WHICH WAS CAPTURED BY CROSS MAGNETIC FIELD

I. I. DEMIDENKO , N. S. LOMINO , V. G. PADALKA , B. G.
SAFRONOV AND K. D. SINELONIKOV

STUDIED THE CONDITIONS OF EMERGENCE AND THE DEVELOPMENTS
OF INSTABILITY IN PLASMA , WHICH WAS CAPTURED BY CROSS
MAGNETIC FIELD AND WHICH PASSED THROUGH DIAPHRAGM .
SHOWED , THAT UPON THE GROUNDING OF METALLIC DIAPHRAGM
OF INSTABILITY PLASMAS ARE NOT DEVELOPED .
PGPH. 1 .
      QUESTION CONCERNING THE STABILITY OF PLASMA .
WHICH WAS CAPTURED BY CROSS MAGNETIC FIELD AND MOVING ALONG
MAGNETIC FIELD        , IS VERY ESSENTIAL UPON
THE CONSIDERATION OF POSSIBILITIES OF UTILIZATION OF THIS
PLASMA FOR FILLING MAGNETIC TRAP .
PGPH. 2 .
      IN WORK       UPON THE INVESTIGATION OF PLASMA .
WHICH WAS CAPTURED BY CROSS MAGNETIC FIELD , WAS NOT
DISCOVERED ANY INSTABILITIES .
PLASMA IN ADDITION TO THAT WAS PASSED THROUGH GAP
DIAPHRAGMS , OR IN GENERAL WAS NOT SUBJECTED TO MECHANICAL
COLLIMATION .
PGPH. 3 .
      IN RIGHT WORK IS REPORTED CONCERNING THE POSSIBILITY
OF DEVELOPMENT OF INSTABILITY IN THIS PLASMA .
PGPH. 4 .
      EXPERIMENTAL APPARATUS FOR THE STUDY
OF THE INTERACTION OF PLASMA WITH CROSS MAGNETIC FIELD
WAS DESCRIBED IN DETAIL IN          .
PLASMA , WHICH WAS CAPTURED BY CROSS MAGNETIC FIELD AND
WHICH IS PROPAGATED IN THE DIRECTION OF THE LINES OF FORCE
, WAS PASSED THROUGH DIAPHRAGM WITH CENTRAL OPENING
BY DIAMETER 15 MM. .
THE DIAPHRAGM INSTALLED ON SUFFICIENTLY GREAT DISTANCE FROM
THE PLACE OF INJECTION OF PLASMA IN MAGNETIC FIELD ( (DIFF)
30 CM. ) ., IN ADDITION TO THAT PLASMA , WHICH PASSED
THROUGH DIAPHRAGM , HAS SUFFICIENTLY THE GREAT RELATION
OF LONGITUDINAL ENERGY TO THE ENERGY OF CROSS MOVEMENT .
ON DISTANCE (DIFF) 100 CM. FROM DIAPHRAGM THE DRIFT
OF PLASMA IN THE DIRECTION OF INJECTION WAS NOT OBSERVED .
PLASMA BUNCHES WERE RECORDED FROM PLASMOSCOPE            ON
DIFFERENT DISTANCES FROM DIAPHRAGM .
```

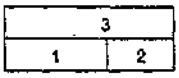
Figure 1 Example of Russian-English machine translation.

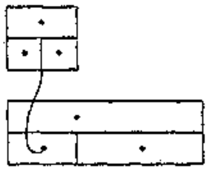
conceived more than ten years ago, and in particular it is the linguistic basis adopted that makes them virtually incapable of improvement. The rapid advances in computer technology provided the impetus for a general reappraisal of machine translation systems and methods. The equipment used in present-day systems has become obsolete, and the arrival of the third generation of computers with higher capabilities of processing speed, core storage, and new direct access devices such as the magnetic disc, made a mere reprogramming of the existing systems unjustifiable both from the economic and scientific points of view. It was therefore decided to devise a new system which would exploit both the new technological resources and past experience in machine translation and linguistic research.

Particular attention is being devoted to the solution of linguistic problems. The main goal here is not to obtain immediately the 'Fully Automatic High Quality Machine Translation' that had been promised in the fifties, but to design a system which will bring about an immediate improvement in translation quality and at the same time permit continuous development after the system becomes operational.

In linguistic analysis, the principal stress was put on the syntax, since this is recognized as the basis for the solution of most of the other problems of translation. The purpose of syntactic analysis is to define the relations between the words of a logical text unit. Since the basic requirement of syntactic analysis is to assign all the words of a sentence into a network of relations according to a set of precise rules, it becomes one of the most powerful methods for the resolution of such problems as the ambiguity of words and the choice between several possible translations of one word. In the design of the syntactic model, allowance was made for the fact that linguistic analysis in machine translation has a purely instrumental function, that is to replace to a certain degree the comprehension of the source text that is necessary for a human translator.

A syntactic relation always consists of three elements: the two terms related to each other and an indication that defines the relation, which may be either explicit, as for instance a preposition or conjunction, implied in one of the terms or in both of them, or simply contained in the word sequence. A syntactic relation, here, is graphically represented as a rectangle divided into three cells where cells 1 and 2 represent the two terms of relation and cell 3 represents the relational element (left).



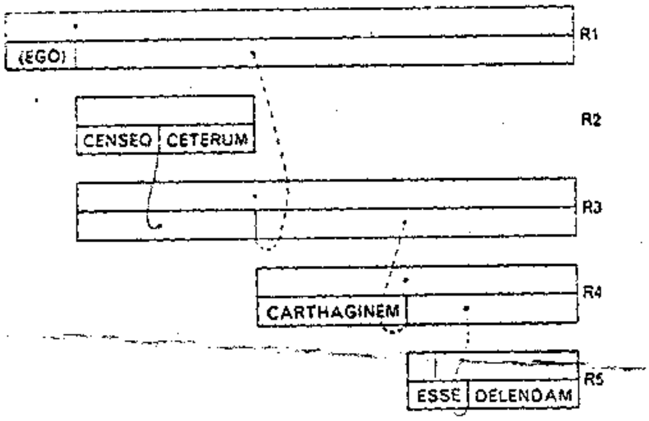


When a relation has been established, the result can be used as an element in other relations. Its function depends on the kind of relation established beforehand. The combination of relations is represented graphically below. The first task of linguistic analysis consists in identifying the relational elements individually, such as all the prepositions, conjunctions, implied elements as in certain word classes (such as the adverb) and word forms (for example, the cases of the noun), and also the purely positional relationships (for example, the noun-noun relation in English; thus 'machine translation' and 'translation machine' differ considerably in meaning). Secondly, for each relational element, or class of elements, the rules are defined that must be satisfied in order to establish a relation (for example, for a preposition as relational element the rule might be that the first term should be a verb before it and the second term a noun after it). The relation is defined as a function of the three elements and their mutual sequence. The purely formal rules as illustrated above are frequently insufficient for the definition of the syntactical structure of the sentence and the subsequent translation.

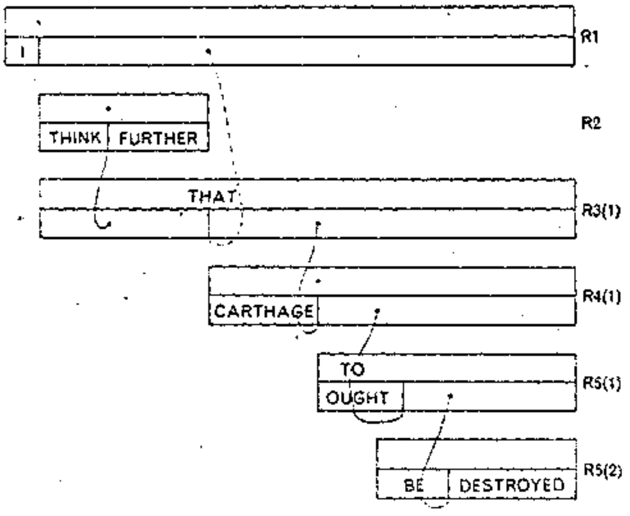
In many cases it is necessary to know more exactly the meaning of the relation, as in the example 'he was arrested by the police' and 'he was arrested by the station'. In both cases the relation can be defined purely formally as 'verb-preposition (by)-noun', but in the first sentence it expresses an agent in a passive construction, while in the second sentence it defines a location. Thus, it may become necessary to split many formally unique relations into a set of different relations as a function of different meanings or translations. To differentiate between the meanings of relations in many cases requires the introduction of new information into the dictionary, as in our example when we need to know whether a certain noun is a potential subject of some activity.

When the syntactical structure of a sentence has been defined, the next step consists in defining the output equivalent of each single relation and the performance, if the need arises, of structural transformations. Suppose that the syntactic structure of the Latin sentence:

CETERUM CENSEO CARTHAGINEM ESSE DELENDAM is:



For a correct transfer into English, this structure must be transformed into:



I FURTHER THINK THAT CARTHAGE OUGHT TO BE DESTROYED.

As the above example shows, the transformations necessary may be of considerable importance, and the information required must often be more detailed than a purely formal identification of the relations. However, although it may be fairly obvious in isolated examples what sort of additional information will be needed, in a large-scale operational system we still do not know what kind of classification of single words and relations will be necessary. Moreover, the existing large machine dictionaries contain almost exclusively the most elementary classifications such as inflection code, word class, and sometimes indications concerning case and preposition government. Therefore, in the first stage of development of the new system, an attempt will be made to obtain optimum results with the information available, and to determine by means of experimental translation runs what additional information is most urgent.

It is expected that the new system will become operational at a very low level of semantic analysis and will produce considerably better results than the present system. Subsequent phases of development will consist almost entirely in a progressive refinement of semantics. It is hoped that progress in methods of automatic classification, such as the theory of clumps [6], will help to resolve some tasks which otherwise might occupy a man's life, as for instance the definition of all potential subjects or objects of all verbs.

Although the development of advanced programming techniques for machine translation can only marginally influence the translation quality itself, the translation speed is vital for the economics of a practical machine translation service. For this reason the portion of the new system concerned with data processing was also designed with a particular attention to performance. In particular, the dictionary-look-up phase, at present the most time-consuming operation in machine translation, was examined most attentively, since the input text must be sorted alphabetically before the dictionary look-up and put back into the original word order afterwards. To avoid these long sortings, special list-processing techniques are used, and it is estimated that the system will work at a speed of some 300 000 words translated per hour, compared with 60 000 words per hour of the present system.

Another important feature of the new system is the SLC-II programming language. SLC, which stands for Simulated Linguistic Computer, is already an integrated component of the Georgetown translation system and was conceived and first implemented by A. F. R. Brown [7]. It is a special-purpose programming language for linguistic applications, and past experience with the Georgetown system has shown that linguistic research benefits considerably from the availability of a symbolic programming language which relieves the user of all data processing and storage considerations and enables him to concentrate on the specific linguistic problems. The new type of SLC language has been completely redesigned so as to make it basically computer-independent, and more flexible, for applications other than machine translation, such as in automatic documentation.

References

[1] Dostert, Leon E. 'The Georgetown-I.B.M, Experiment' in 'Machine Translation of Languages'. The Technology Press of M.I.T., Cambridge (Mass.). 1957.

[2] Chomsky, N. 'Syntactic Structures'. Mouton, The Hague. 1957.

[3] Ceccato, S. (Ed.). 'Linguistic Analysis and Programming for Mechanical Translation'. Gordon and Breach, New York. 1961.

[4] Perschke, S. 'The Use of the 'SLC' System in Automatic Indexing'. In *Proc. F.I.D./I.F.I.P.* Conf. Rome. 1967. (In press.)

[5] 'Language and Machines: Computers in Translation and Linguistics'. Automatic Language Processing Advisory Committee. 124 pp. Nat. Acad. Sciences. Washington. 1966,

[6] Needham, R. M. 'The Theory of Clumps'. Cambridge Language Research Unit, ML 139. 1961.

[7] Brown, A. F. R. 'The "SLC" Programming Language and System for Machine Translation'. Report EUR 2418. e. Presses Academiques Europeennes, Brussels. 1965.