MACHINE TRANSLATION:

MOSCOW INTERNATIONAL SEMINAR

I. I. OUBINE

Head, Machine Translation Department
All-Union Centre for Translation of Scientific
and Technical Literature and Documentation
Ul. Krzhizhanovskogo 14, korp. 1
117218 Moscow B-218

An International Seminar on Machine Translation took place in Moscow from 25 to 27 November, 1975. About 200 scientists from People's Republic of Bulgaria, the German Democratic Republic, Czechoslovak Socialist Republic, People's Republic of Poland, and the Soviet Union participated in the Seminar which was organized by the All-Union Center for Translation of Scientific and Technical Literature and Documentation.

In his opening speech the head of the All-Union Center for Translation (ACT), Dr. V. N. Gerasimov, said that the expansion of interational contacts and internationalization of science had promoted the growth of the translation activities in the USSR. In the country at present translation is being done by various specialized and departmental organizations. In 1975, for example, the ACT translated more than 30,000 author's sheets of scientific and technical literature and documentation. In the nearest future ACT alone will reach the volume of 50-80,000 printer's sheets a year. Dr. Gerasimov sees the way out of this situation in the speediest possible development and installation of industrial systems of machine translation. This was also the conclusion of the Temporary Scientific and Technical Commission on Machine Translation of the State Commission on Machine Translation of the State Committee on Science and Technology of the Council of Ministers of the USSR which worked in 1972-1973

At the plenary session, four reports dealing with general principles of construction of machine translation systems and ways of improvement of MT quality were delivered. Yu. N. Marchuk (ACT, Moscow) draws our attention to the increased role of computer dictionaries in automatic data processing systems and namely in MT systems. Since the quality of MT largely depends on the word-list and volume of the dictionary and the information contained in the dictionary entries, these dictionaries must, according to Marchuk, be compiled with due regard to distributional and statistical methods. It is important to make wide use of contextual information. Inter- and post-editing are absolutely indispensable in running the first industrial MT systems.

This idea is developed in a joint report by B. D. Tikhomirov, Yu. N. Marchuk, and I. I. Oubine (ACT, Moscow), who put forward a new approach to inter- and post-editing for correction of input and translation errors. Interediting follows a search in the dictionary, automatic analysis of new words and formation of the information required for future processing. The intereditor either confirms or corrects the grammatical information of new words and can also correct input errors, supply translations for new words, etc. At the intereditor's command the system accumulates new words for the posteditor.

The posteditor exercises phrase-by-phrase control of the translation. If the translation of a phrase does not satisfy him he edits it with the help of a display unit and, if necessary, sends the original phrase and its machine translation to a special storage device for further analysis aimed at the improvement of the MT algorithm.

In his summarizing report, Dr. A. Ljudskanov (PRB) singled out the main stages of MT development, and showed what an unfavorable effect the ALPAC report and the pessimistic statements of Dr. Y. Bar-Hillel had produced on MT. MT systems aimed to operate on an industrial scale in the nearest future are to be developed on the basis of the "selective" strategy, i.e. they are to use only that information "which is necessary and sufficient for the given aim for the given pair of languages." Dwelling upon this thought, Dr. Ljudskanov said in his second report, "Lexeme dictionary for MT systems", that in the MT system which is under development in Bulgaria "deep" difficulties are shifted to "surface" levels (lexical and morphological) and solved with the help of context analysis.

The fourth report at the plenary session was read by Dr. R. G. Piotrovsky, who thinks that for the time being those who develop MT systems must not strive to achieve 100% efficiency of text processing and these MT systems must be based not on deductive generative grammar but on deductive-inductive linguistics of text. According to the linguistics of text, MT systems should be developed consecutively from simple algorithms operating with units of the surface levels of the text system to more complicated algorithms oriented to the deep levels of the text systems. Dr. Piotrovsky considers that vocabulary contains the lion's share of syntactic and semantic information of the text and consequently the basis of any MT system must be a bilingual automatic dictionary with the output word list of the thesaurus type: Such MT systems can be developed within reasonably short periods of time and will, according to Dr. Piotrovsky, enable the consumer to derive all semantic information from the input text.

Thus the main speakers put forward as the cardinal task of the present day the creation of MT systems not yet oriented to high-quality translation but working industrially

More than 60 reports were made at the seminar which were distributed into four sections: (a) computer dictionaries; (b) automatic analysis and synthesis of texts; (c) semantic analysis of texts; (d) mathematical and program maintenance.

Various trends in the theory and practice of machine translation were represented at the section on computer dictionaries. A number of reports were read by members of the Speech Statistics group (the work is under the scientific guidance of Dr. R. G. Piotrovsky). Scientists of this group compile dictionaries for MT within the framework of one scheme in accordance with the postulates of the linguistics of text; their MT dictionaries consist of two parts: a dictionary of commonly used words and changeable terminological dictionaries for various fields of science and technology. The input entries of these dictionaries are either single word forms or combinations of several word forms. Morphological and syntactic information is written straight in the dictionary entries. The volume of syntactic information is relatively small. This information includes indices of grammatical word-class, transitivity-intransitivity for verbs, type of government and some other syntactic characteristics. Dictionaries of input and output languages have charts for coding information and correspondence charts which in this MT system are the principal linguistic algorithm. All programs of automatic processing of linguistic information are based on these charts. The authors of these reports pay great attention to methods of coding information and compressing codes in the computer. The computer realization of such dictionaries makes it possible to get an interlinear version which is not actually a translation but gives the consumer the main contents of the original text. During the discussion, many participants of the seminar, criticizing this approach to MT, pointed out that automatic dictionaries should not be compiled separately but as component parts of MT systems. The coding of lexical information in an MT system should be preceded by the elaboration of blocks of grammatical analysis and synthesis which use lexical information and in their turn may affect its composition and coding.

A different approach to principles for development of MT systems was presented by Z. M. Shaljapina in "The ARAT system: dictionary, grammar, and their use in automatic analysis" (Moscow, State Institute of Foreign Languages). The central component of this Anglo-Russian Automatic Translation system (ARAT) is a dictionary of a special type (Anglo-Russian Multiaspect Automatic Dictionary--ARMAD) in which every entry gets complete and diverse characteristic properties of its linguistic behavior on all levels of description: morphological, lexical, syntactic, and semantic. For formal recording of all required information, ARMAD employs syntactic patterns, lexical functions, demands imposed by the word on the linear and structural context, rules of standard and individual modifications of lexical and syntactic structures with this word, formal semantic representation, semantic selection of syntactic valencies of the key word, etc.

The abundance and variety of information contained in the dictionary entry of practically every word is the principal advantage of ARMAD as compared with other types of automatic dictionaries. Wide use of syntactic and expecially semartic information makes it possible with the help of this dictionary in the framework of the ARAT system to solve such a cardinal problem of MT as grammatical and lexical ambiguities. Data on lexical cooccurrence of the original English word and its Russian equivalent as well as on required syntactic transformations, it is hoped, will ensure an exact and idiomatic translation of the original text But at the same time the abundance and variety of the information in ARMAD naturally leads to a more complicated entry structure and multiplies the difficulties which linguists have to overcome when writing entries. At present the difficulties and possibilities of algorithmic realization of full-scale dictionaries of this type have not been investigated.

A number of reports were devoted to compiling various types of auxiliary dictionaries and creation of whole systems for automatic lexicographic work. In their reports, Yu. N. Marchuk, N. G. Tikhonova, and I. I. Oubine (all of ACT, Moscow) informed the participants of the seminar of the work now being done to develop a bilingual automatic lexicographic system to compile frequency and semantic frequency dictionaries and bilingual concordances as auxiliary material for MT dictionaries.

E. V. Vertel and V. A. Vertel (State Institute of Foreign Languages, Minsk) submitted a joint report on the elaboration of an algorithm and a set of programs to compile a frequency-alphabetic dictionary and concordance on a medium-size computer. The set of programs is designed to process an original text of up to 300,000 words.

An interesting report on the compilation of a reverse French dictionary was submitted by E. L. Kozmina (Applied Mathematics Institute, Moscow), who worked out a method of getting a wordform by its ultima or by one wordform representing a whole class of words inflecting uniformly. This method makes it possible to reduce a reverse dictionary of word-forms to the size of an ordinary dictionary.

In this session the participants in the Seminar discussed the principles of compilation of MT dictionaries and came to the conclusion that it is necessary:

- to consider an MT system as a whole in which all parts are interlocked, while the dictionary is its principal component;

- to design MT systems for sublanguages and consequently compile automatic dictionaries for well-defined topics;

to pay special attention to strict selection of entries for automatic dictionaries, and for this purpose compile various auxiliary dictionaries and concordances;

- to increase the volume of diverse syntactic, semantic, and lexical information in automatic dictionaries;

- to attach great importance to the structure of automatic dictionaries. According to a number of the participants of the seminar, the optimum structure is a combination of a dictionary of stems and a dictionary of wordforms. The dictionary of word-forms contains more frequent words and the dictionary of stems contains the less frequent words.

The section on automatic analysis and synthesis of texts attracted the largest number of reports, reflecting the intense interest of linguists in these important parts of MT systems. A series of interconnected reports was read by scientific workers of the Computational Center of Leningrad State University Dr.

G. S. Tseitin proposes to use for automatic syntactic analysis models which do not simply contain a set of rules for construction (or "filtering") of admissible syntactic structures but also mark out among these rules more or less "preferable" ("normal", "nuclear", "productive"). B. M. Leikina (Leningrad State University) is working on a model of an English grammar which permits under certain limitations nonprojectivity while fulfilling the projectivity demands for the majority of cases. At present this group is working on a more complicated model which takes into account among other things the order of establishment of syntactic links in a particular structure' A number of experiments have been carried out with the first variant of this grammar. These experiments were aimed at checking the presence of the correct analysis and absence of a fixed incorrect analysis and were carried out under the conditions of man-machine interaction during which a part of the intermediate results were rejected by the man. Complete analysis has been tested only on several sentences.

The principles of analysis in the ARAT system were expounded in this section. Besides the report of Z. M. Shaljapina already mentioned, three reports on less general problems were read: L. A. Afonasjeva, "Disambiguation of homonymy in the ARAT system", T. N. Nikanorova, "Prepositions in the ARAT system", and O. A. Sternova, On a model of Russian inflexion in the ARAT system" (all State Institute of Foreign Languages, Moscow). In this MT system the disambiguation of homonymy is not singled out into a separate block but is carried out during the analysis simultaneously with the fulfillment of other tasks. Prepositions in the ARAT system are treated like any other lexical unit of the text. Meaningless prepositions, i.e. prepositions serving as surface-syntactic indications of strong government and having no influence on the meaning of the text are eliminated during the transition from combined syntactic structure (CSC) to semantic representation. Meaningful prepositions, i.e. prepositions having their own meaning, are elements of the semantic structure and are preserved during the transition from CSC to semantic representation.

E. E. Lovtsky (ACT, Moscow) proposed formal means for the description of the syntax of natural languages. The description is an oriented graph, in the nodes of which are symbols of syntactic classes and of subgrap The information on the dependency structure of strings is recorded on the edges of the graph. The description of a natural language syntax, made with the help of the suggested formal means, can be used in a system of automatic syntactic analysis. For the analysis of a phrase, it is necessary to find in the graph all occurrences of a string of syntactic classes corresponding to this phrase and to read from the edges of the graph the information about the structure of the string. This procedure is done automatically by a special algorithm. As a result of the analysis we get immediate constituent trees and dependency trees of the analyzed phrase. In case several trees correspond to the analyzed phrase then the choice of the tree corresponding to the lexical units of the phrase is made on the basis of the information contained in the automatic dictionary. The said description is being done for the analysis of English scientific and technical texts.

The joint report of E. Benesova, A. Bemova, S. Machova, and J. Panevova (Czechoslovakia) contains characteristics of cardinal components and principles of functional generative description designed to get semantic representations which in their turn are the basis for synthesis of sentences of a natural language.

Other reports at the section on automatic analysis and synthesis of texts were of a less general character. Dr. Klimonov (GDR) presented a set of criteria for automatic identification of antecedents of personal and possessive pronouns in Russian and German. Dr. Starke (GDR) read a report on a method of transformation of Russian structures with German equivalents having different dependency representations. Each transformation is performed by means of elementary operations: insertion, deletion, alteration of a node, etc.

The reports submitted to the section on semantic analysis touched on some narrow problems. A number of reports show their author's efforts to solve their problems within the framework of automatic text processing. These reports are the definition of the concept of answer in question-answer structures of texts (Dr. Konrad, GDR); semantics of prepositions (Leontjeva, ACT, Moscow) ard Nikitina (Institute of Linguistics, Academy of

Sciences, Moscow); semantic and syntactic analysis of Russian words with the meaning "quantity" (Iljin and Smirnova, Leningrad State University); communicative structure of the English sentence (Korolev, ACT, Moscow); semantic analysis of headings of Japanese patents (Zeinalová, .Samarina, and Shevenko, Institute of Oriental Studies, Academy of Sciences, Moscow).

An interesting report was submitted by N. A. Kuzemskaya and Dr. E. F. Skorokhodko (Institute of Cybernetics, Kiev), who reported work on quantitative criteria for estimation of translation quality. They propose to solve this problem with the help of a semantic language which would meet the following conditions: (a) ability to express the required information; (b) possibility of getting the necessary quantitative estimates; (c) possibility of translating into this semantic language from the original language and from the language of translation.

Semantic networks can be used as such a semantic language. Comparison of two speech semantic networks--the primary corresponding to the original text and the secondary to the text of the translation--makes it possible to define a number of parameters characterizing various aspects of the quality of the translation. The main aspects of translation quality are completeness and accuracy. Completeness is calculated from the proportion of the primary network contained in the second, and accuracy from the proportion of the secondary that coincides with the primary.

The problem of automatic recognition of key words in text was the main concern of the reports of R. A. Kovalevitch, V. A. Sorkina (State Institute of Foreign Languages, Minsk), and G. S Osipov, A. I. Chaplja (Mahachkala State University). Kovalevitch and Sorkina singled out more than 50 elementary semantic units and, relying on formal characteristics of the text and using distributional statistical methods, they singled out and formalized rules of combination of elementary semantic units in free word-groups. The result of the analysis is a matrix of relations of components in English word groups and their Russian translations with all necessary grammatical information. The set of formal characteristics of all components of a free English word group determines the corresponding Russian equivalent. In their report, Osipov and Chaplja spoke about an attempt to construct (by the thesaurus method) formal means useful for elaboration of an algorithm of recognition of key words in a text. These authors build this formal apparatus on the basis of estimating the degree of synonymy of words in the structures of the input text.

To our regret it should be admitted that in spite of the universally recognized importance of the semantic component for MT, the seminar lacked reports in which authors put forward methods and instruments of formal representation of word meanings for the solution of problems of analysis and synthesis of texts during MT. The participants of the seminar did not pay sufficient attention to the search for optimum ways of elaborating formal semantic languages for MT.

In the software section, ten reports were made, three of which are characterized by a broad approach to this important part of MT systems. These are the reports of N. A. Krupko and G. S. Tseitin (Leningrad State University), V. S. Krisevitch and I. V. Sovpel (Institute of Foreign Languages, Minsk), and D. M. Skitnevsky (Institute of Foreign Languages, Irkutsk) The joint report by Krupko and Tseitin sums up the research work carried out during the last 13 years, involving development of software to be used in computer experiments on MT. At present the software of the MT model at the MT laboratory of Leningrad State University comprises the following components: (a) the system of·symbolic representation of the linguistic content of the model; (b) the machine representation of the linguistic content; (c) the compiler for translating the symbolic representation into the machine representation; (d) the interpreter which employs the machine representation to process a given text in accordance with the aim of the experiment. Such a structure of the software is motivated mostly by the necessity of changing the linguistic content in the course of experiments and by the impossiblity of keeping in core memory more complicated linguistic information in the form of a compiled program. The same approach with an appropriate shift in the structural role of each of the component parts seems, in the opinion of the authors of the report, to be quite reasonable in developing practical MT systems.

The report by Skitnevsky considers some basic principles of software development for linguistic investigations. The main idea of Skitnevsky's model is to consider as interdependent the following three levels of software: (a) the conceptual level--a model of linguistic process; (b) the compositional level--an input language; (c) the programming level--the implementation system. The development of software includes the following five stages: the first stage is an analysis of a set of algorithms which are given in their verbal description (i.e. as flow charts). The analysis results in an informal definition of the class of algorithms. The second stage consists in specifying this formal definition. As a result we have a logical and mathematical framework, which is a sufficiently precise and dynamic description of basic concepts occurring in the informal definition (MODEL). At the third stage a data bank and a service program package using the standard software are formed on the basis of the set of objects and operations of, the model. The fourth stage consists of working out an input language which permits the user to state his tasks in terms of the MODEL. The fifth stage designs an operational system which uses the standard software and remains intact in respect to relatively extended MODEL, its data bank and service program package. The MODEL built along these lines and an input language form a convenient metalanguage for the description of the computer semantics of natural languages.

The report by Krisevitch and Sovpel emphasizes that an MT system operates with large information files including input and output automatic dictionaries containing grammatical and semantic information; therefore the efficiency of an MT system depends to a great extent on the degree of optimization of information processing. The authors believe that the efficiency of use of storage and the minimization of average access time are determined by such factors as the nature of information, the request frequency, and the properties of storage devices. In keeping with these principles, the authors constructed their model of the information base for an MT system. The other reports discuss less general problems of software relating to automatic text processing. The report by N. G. Arsentjeva (ACT, Moscow) describes the results of the machine experiment which was aimed at forming surface syntactical characteristics of a word on the basis of its semantic description. The input information is a semantic description of some word, represented as a formula containing semantic elements and operation signs. The author has devised an algorithm which supplies the linguist with a list of possible realizations in the text of the syntactic pattern of the processed word on the basis of the semantic description of the word, its syntactic pattern and some other linguistic information. The program is written in the algorithmic language of recursive functions. The experiment shows that the said language is quite applicable for linguistic purposes.

S. A. Ananjevsky and P I Serdukov (Institute of Cybernetics, Kiev) informed the seminar of their work on the development of a system of automatic syntactic analysis of text for purposes of automatic abstracting, information retrieval, and MT. The system permits one to obtain the following characteristics of the English verb: search area (zone of syntactic dependents of a verb in a sentence), position of a verb (predicate), which is the initial point of the search; the immediate distribution of the verb; syntactical compatibility of the verb; syntactic compatibility of the dependents (string of verb dependents). This system is implemented on the Minsk-32 and permits one to obtain a complete formal description of syntactical dependencies of the English verb.

The report by L. N. Beljaeva and E. M. Lukjanova (Gertsen Pedagogical Institute, Leningrad) formulates principles of compiling an automatic polytechnical dictionary for MT. The dictionary is a system of sets of linguistic data as well as algorithmic, linguistic, and programming units of a Russian text, different for different linguistic algorithms. The structure of such a polytechnical dictionary gives the possibility of organizing this system as a linguistic data bank. E. V. Krukova and T. I. Gushchina reported the results of the application of the PL/I language for an automatic analysis of texts. The authors have written a number of programs in PL/I, designed to compile frequency and reverse dictionaries, for morphological analysis of Russian words, and for resolution of ambiguity of French nouns. In their opinion, PL/I is well suited to the purposes of automatic text processing.

Summing it up, it should be noted that the reports submitted to this section varied considerably from the point of view of the problems discussed and the methods employed due to lack of uniformity in using programming languages and types of computers

The main characteristic feature of the seminar, in our opinion, is an emphasis on developing practical systems of MT for well-defined sublanguages of natural languages rather than devising global language models of which MT is only a component. The most essential thing now is not to carry out experiments but to develop practical systems of MT operating on a large scale.

At the concluding session the participants of the seminar adopted a memorandum which admitted the importance of the seminar for the development of practical MT systems and appealed to ACT and the International Centre of Scientific and Technical Information to hold another seminar in Moscow in 1977. The participants of the seminar think it necessary to promote further coordination and integration of research in MT in the USSR as well as in the framework of the International System of Scientific and Technical Information. They find it reasonable to determine the main body responsible for problems and to entrust AGT with this important task.