Machine Translation in the USSR

YU. N. MARCHUK

All-Union Centre
for Translation of Scientific
and Technical Literature
and Documentation,
ul. Krzhizhanovskogo, 14,
117218 Moscow. USSR

The state-of-the-art in machine translation (MT) in the USSR is considered in three aspects: theoretical, practical, and informational. A survey is presented of theoretical studies in the field of linguistics for MT purposes. Then a description is given of several operational MT systems like AMPAR, NERPA, and FRAP. The question of MT product use in information services is discussed with a view of assessing its economic efficiency.

The machine output quality of modern machine translation systems is connected inseparably with the nature of translation models in use. One of the main trends in simulation of translation is research study and reproduction of actions performed by a human translator.

Existing translation models involve all operations that are specific to human translation. On the other hand, all actions of a human translator involve operations that can, to a certain extent, be simulated in computer-aided translation.

When simulating translator's actions, the authors rarely build a general model, but rather confine themselves to a certain type of translation. Thus, Z. M. Shalyapina [1] considers written translation as a continuity of operations, part of which can be easily implemented by the computer (operations on the surface level), and part of which is difficult for machine implementation. A. F. Shiryaev [2] offers description of a model of a simultaneous translation functional system based on theoretical and experimental studies of a simultaneous translator's actions. The simultaneous translation functional system is treated by the author not as a system in general, but as a specific system on the assumption that simultaneous translation can be mastered normally by means of development and rearrangement implemented in functional systems of other kinds of translation. Basic techniques of simultaneous translation are: timing, understanding of a source text, search for and implementation of translation options, and their verification and correction. A prime role is assigned to the timing technique represented by various levels of actual cognition, unconscious verification, conscious verification, etc. Yu. N. Marchuk does not think that simultaneous translation is absolutely unique as a form of translation and in his concept of the translator's actions (geared mainly to simultaneous interpretation), he docs not stipulate the specific features of the latter but links up the interpreter's actions with a certain concept of linguistic understanding [3]. The increased interest in interpreter's actions and simulation thereof corresponds to a similar trend elsewhere in the world today and reflects the importance of the 'transfer stage', i. e. the translation correspondences proper in constructing modern systems of machine translation.

Another important direction of the theory is a study of specific features of sublanguages in connection with simulation of translation. It has become apparent after a period of experimental use of many MT systems that the quality of translation can be improved if the specific features of sublanguages which aid automated analysis are reflected in dictionaries and algorithms of analysis and synthesis. The theory of sublanguages or language subsystems was first formulated by the Soviet scientist N. D. Andreev [4]. In the recently published book by L. L. Nelyubin [5], a theory concerning sublanguages from the viewpoint of machine analysis and translation is offered. The sublanguage is described by four models: functional-communicative, statistical, informational, and linguistic. On the basis of these models, a translation system is being constructed for organisational and management documents translated from English into Russian. This system is based on a computer dictionary compiled specially for this purpose.

Problems of lexical analysis based on formal indices, even if these are not explicitly tied up to MT, are of great importance to it since the compilation of dictionaries is the most labor-consuming part of any MT, whereas their completeness and adequacy to the formulated objective greatly facilitate the improvement of the MT quality. First of all, in vocabulary analysis, attention is focused on the ambiguous nature of a word (polysemy), whose resolution is rather important for translation. N. Ya. Serdobintsev [6] refutes the assertion of the outstanding Soviet linguist R. A. Budagov [7] that polysemantic words comprise about 80 per cent of any glossary. He gives data certifying that out of 10,515 words analysed in two volumes of the 'Modern Russian Literary Dictionary' there are 8,657 monosemantic words, or 83.5 per cent, and 1,872 polysemantic ones, or 16.5 per cent. Thus the real picture is directly opposite to that obtained on the basis of an intuitive notion. The concept of word ambiguity in MT is associated dm with traditional polysemy but rather with the existence of several translation options; nevertheless this data is undoubtedly of a great interest.

Modern MT is geared to translation of scientific and technical texts. A complex of problems associated with the correlation between terminological and common use vocabularies is under intensive studies. In particular, how words of common use vocabularies become terms and vice versa, how terms come into common use, is of a great interest [8]. Most authors conclude that a word's meaning is clear only from the context.

Comparative language study has recently become a special tract in linguistics defined as contrastive linguistics. This linguistics always deals with data of two languages or of several languages studied in pairs.

The immediate object of contrastive linguistics is bilingual systems or subsystems, and its problems are closely interwoven with the problems of translation [9].

Studies of language semantic level are under way in connection with MT. Results indicate that one should not overestimate the possibilities of the semantic level. If a sentence has passed through all levels of analysis and obtained a unique semantic representation, then synthesis will ensure a normal translation for it. However, this level is incapable of resolving all ambiguities, on the contrary, it can make a syntactically monosemantic sentence ambiguous [10]. Efforts are being made to employ a semantic dictionary of the combinatorial and glossary type in the formal model of language and to resolve the polysemy in machine translation [11]. A tendency to formalise the meaning of word chains exceeding a word or sentence in length leads to the concept of developing text meaning as a whole. Texts written in natural languages do not accept fixed semantics. The meaning of a word can be determined only in a definite context. A word as such, without any potential context, is simply a sign or the name of some object, but has no meaning since it is not the element of a meaningful statement. Each word harbors a multitude of meaningful texts in which it can be included. Not just any set of words is meaningful for the simple reason that for some words in it the remaining words cannot form an appropriate context. Naturally, the meaning of the text can become clear to a reader only to that extent to which he is familiar with the language, i. e. to which he knows the potential contexts of each word. Thus, the question concerning the nature of the meaning which a word has can be confined to the question on the nature of the meaningful statement [12], The authors describe the concept of statement meaning through the concept of role structures, which are understood to be an abstraction of a functionally integral situation as a set of 'roles' regardless of the particularities of its elements.

Everything mentioned above illustrates the multilateral policy in theoretical studies in the field of MT, which covers as before a wide range of problems.

The practical activity in MT is based on operation of a number of MT systems with post-editing. In the Ail-Union Centre for Translation of Scientific and Technical Literature and Documentation, the AMPAR machine translation system (Automated Machine Translation from English into Russian) has been operating for a number of years [13. It is intended for translation with post-editing of texts covering radio-electronics, computer science, programming and a number of other technical fields.

The linguistic support of the AMPAR system is based on a special translation model using translation correspondences and consists of two components: the dictionary and grammar ones. The entire translation process is divided into 17 stages, each performing a specific analysis, translation or text synthesis operation. Source text analysis (stages 1—7) covers morphological analysis and word form matching against the dictionary, search in the text, analysis and translation of set expressions, resolution of all forms of homonymy, syntactical analysis by parts of speech and by parts of the sentence, Translation *per se* (the set of 'transfer' stages), stages 8—14, involves translation of unambiguous and ambiguous words using the contextual environment analysis. Synthesis of the target text (stages 15—16) is performed in two stages: syntactical synthesis, i. e. establishment of syntactical and morphological correspondences between the English text and the Russian one, and morphological synthesis. The entire translation process is completed with listing of the target text (stage 17). Listing can be accomplished using various options: page-by-page listing, parallel listing of the Russian and English text, etc.

The dictionary component of the system represents a sophisticated interaction of a number of dictionaries. The source dictionary (over 25 thousand words) is subdivided according to subject. The following English dictionaries are in operation: common vocabulary dictionary, the general technical vocabulary dictionary, and the computer science and programming dictionary. The word combination dictionaries are also subdivided according to subject fields. The target Russian language dictionary contains about 35 thousand dictionary entries. Translation is performed with the aid of tables matching unambiguous English words with Russian equivalents. The tables correspond to subject division of the source dictionary. Ambiguous words are translated using specific algorithms that establish this or that translation of a word by analysing its context usage [14, 15]. Updating of available dictionaries and creation of new dictionaries is a routine process.

The translated text goes to post-editing and is then delivered to the customer as, per his request, a typewritten copy, the line printer listing, or on magnetic tape. The quality of translation is such that it can be understood by a specialist.. This allows us to deliver unedited text to the customer in most cases as preliminary information to speed up its use. The volume of literature being translated totals several hundred signatures per year, and the number of new orders is constantly increasing.

Translation of huge volumes of texts under industrial conditions serves as a good updating source and a system dictionary enlargement source. Work is being done on describing new sublanguages to extend the subject fields of texts translated. Practice has shown that to satisfactorily cover the texts of a sublanguage, it is sufficient to supplement the system dictionary by 4—5 thousands of Russian and English lexical units and the word combination dictionary by 5—6 thousands of dictionary entries. Eight scientific research workers can manage this job in 3—4 months.

Aside from the quantitative enlargement of the dictionary files, routine work is carried out on qualitative improvement and upgrading of the system. AMPAR-2 is being created in which the entries of source dictionary will be based on a wider use of semantic and word combination properties of a word. The system also provides for the perfection of syntactical analysis, thereby allowing translation quality to be enhanced. The idea used as the foundation for the AMPAR system support and implemented in a model of translation correspondences ensures perfection of the system without any dramatic changes of its framework and in such a manner that additions do not deteriorate system performance, as it sometimes occurs with systems in operation.

The NERPA translation system (Automated Machine Translation from German into Russian) employs the same linguistic and programming principles used by the AMPAR system. The specific features of German grammar have been incorporated into algorithms of this system [16). In particular, the morphological analysis based on a relatively broad system of inflections plays a considerable role in the system. At the same time a broad homonymy of inflections required the development of a special index system permitting the homonymy of stems to be avoided.

The NERPA system features the morphological (word formative) analysis of the words that have not been found in the dictionary. Such words are then synthesised in Russian in the form of an artificial word with a regular suffix attached to the stem available in the dictionary (исаскание — *search,* покрывание — *coverage).* The NERPA system differs from the AMPAR system mainly in the analysis of composite words so typical for the German language. The general principle is disintegration of the composite words into the component stems followed by their synthesis in the form of a Russian word combination. For example, *Informationsverarbeitung,* is translated as обрабока информации *(data processing).* The composite word analysis reduces significantly the volume of the German and Russian dictionaries. Instead of 50—60 thousand words, were sophisticated analysis not available, the dictionary contains 10—15 thousand words. Since the number of composite words in German texts is virtually unlimited, the composite analysis stage is rather important for the system and represents its characteristic feature.

As compared to the AMPAR system, the NERPA system widely uses semantic and syntactical codes which are employed in part due to the necessity of distinguishing syntactical homonymy but mostly due to the increased number of semantic classes. For instance, nouns may fall into classes denoting space, animals, organisations, artifacts, qualities, processes, etc. The differentiation of semantic classes facilitates selection of a Russian equivalent for the ambiguous German words. Thus, the German word *Seite* will be translated as страница *(page)* if accompanied by words with semantics of a number, or otherwise as сторона *(side). Ausstellung* is translated as установка *(installation)* provided that the neighbouring word has the semantics of the artifact, otherwise it is translated as составление *(putting together)*.

The NERPA system has been recently put into experimental industrial operation. The number of translated texts is small. Efforts are currently being made to enlarge the dictionary, update the files and expand the subject fields. The main engineering field of system application is programming and computer science.

Both the AMPAR and NERPA systems have unified software. Software features the following [17]:

— division of the translation process into a number of stages;

— subdivision of processing files (dictionaries, schemes, tables) into subject field subfiles;

— use of a specialised programming language alongside with the Joint Computer System (ES) Assembler;

— use of a language support (Process Control Language) to specify input/output instructions for files handled and modes of handling;

— the opportunity to reorganise the system structure (creation of various versions to select the most efficient system version);

— capabilities of obtaining the results of system operation at any stage in the form convenient for analysis in the verification mode.

A modular structure concept has been employed in the software system complex whenever individual problems are solved by stage programs, each consisting of program modules. The modular structure concept also pertains to information files (dictionaries, tables, ect.).

Since modules are relatively independent, it is possible to modify programs and information files in a comparatively simple manner by developing and including new modules or changing the sequence of their operation.

Great attention has been paid throughout system development and operation to questions involving the linguist's efficiency in handling the system. As a result, the system's linguist can:

— directly participate in creation and debugging of the programs (schemes) that implement the specific algorithms for processing compound word combinations, translation of ambiguous words, resolution of homonymy and analysis, i. e. participate in those stages which are most likely to change when system capabilities are expanded (a specialised language has been developed to simplify the process of programming and updating);

— obtain information about the words which are not available in the system dictionaries and about the typical errors that occur in the translation process;

— quickly localise an error when translating and determine its nature (selective printout of the system operation results is employed at any assigned section of the text providing highly detailed information with the accuracy reflecting functions of an individual scheme operand);

— without hindering system operation, create various versions of the system. Each version can include new and/or modified schemes or a modified order of their operation;

— verify operation of the created versions using a wide spectrum of texts to select the most efficient version and to include it into the work file as a work version;

— trace the state of information files.

In both the AMPAR and NERPA systems, the operator communicates with the work and service routines in the process of their functioning and sets their operation modes by means of instructions written in a special process control language. This language contains a set of directives. Each directive causes the module to perform I this or that operation.

The NERPA program complex differs from that of the AMPAR system in routine and information modules developed additionally to take into account the specific nature of translation from German into Russian.

The FRAP machine translation system (Automated Machine Translation from French into Russian) operates using somewhat different principles [10, 18] whose essence lies in (he explicit use of the semantic level and in producing semantic translation with validity checks on a contextual level but not on that of translated correspondences, In the first version of the system (1976—1980), the main attention was focused on development of the linguistic support: linguistic structures of various levels — morphological, syntactical, semantical, grammar and algorithmic complexes. The existing version of the system proves the validity of the chosen linguistic ideology that, at any given moment of translation, ensures availability of information on all levels which have been built by that moment.

The software development immediately follows the linguistic support.

The linguistic support is not adjusted to a particular subject matter. The main dictionaries arc formed on the basis of the unspecialised common vocabulary. The terminological dictionary covers three subject matters electronics, computer science, aviation and aircraft construction.

The FRAP system has a modular structure. Its largest subdivision is the subdivision into dictionaries and grammars or, to be more precise, into a dictionary complex and a grammatical and algorithmic complex. Each complex extends to the following levels:

— analysis: graphemic, morphological, syntactical and semantic;

— translation of significant lexemes, relational words, syntactical links, grammatical classes, pronouns;

— synthesis: semantic, syntactical and morphological.

The system operates in several modes. The first one is auxiliary word-by-word translation. This mode enables us to check the main system dictionaries, the source French morphological and syntactical and the target Russian morphological and grammatical dictionaries.

The second one, the principal mode, includes the syntactical component which references the semantic component to verify the meaning of links and translated equivalents. Translation itself is done through syntactical representation of individual sentences, which is why the mode is called a syntactical one. The third mode is a semantic or textual-and-semantic one; it has been installed in the current system version but can be implemented only in the next system version. This is translation through semantic representation, which may be accompanied by conciseness and semantic editing of the text content. Lastly, the fourth mode is an informational one, which assumes selective dissemination of information to the customer. The system must ensure translation of only those text extracts which meet the customer's informational requirements.

In the current FRAP version, interface between the syntactical and semantic components has been worked out. Thus, a sentence is described in terms of two representations: a syntactical one and a semantic one. These representations interact as follows:

(a) realisation in the semantic representation only of those dictionary-covered word meanings which correspond to the given syntactical representation;

(b) rejection of certain doubtful links found in the syntactical representation on the basis of semantic representation.

Much attention is being given to perfecting the software so as to achieve the flexibility required for adequate simulation of operations concerning structure transformations in the course of machine translation. Programming and debugging of an entire cycle of syntactical analysis using the PL/1 language proved to be highly labor-consuming and practically non-agreeing processes. A decision has been made to change over to a more dynamic programming language in which program development and debugging can be performed by the linguists themselves. This language is a variant of the standard statement language developed for the AMPAR system.

In the FRAP system, four different machine data representations are employed as follows:

— pre-syntactical level representation which uses information in the simplest form;

— most consistent and system-organised representation for the syntactical stage, which is phrase-oriented;

— text-oriented representation for the semantic stage. It slightly differs from the previous one in that it has a larger depth due to semantic information;

— representation for synthesis. It is phrase-oriented and word-form oriented and can be reduced to the second and third representations.

It is thought that a more convenient detection of translation units will be achieved in the FRAP system and that this will make for improved translation quality.

In the Ail-Union Centre for Translation, machine translation systems are being worked out in parallel with work on an automated dictionary designed to assist a human translator and editor. At the moment, this dictionary contains English, German, French, and Hungarian lexical files and is oriented to computer science and aviation fields.

The Chimkent Pedagogical Institute provides lexical industrial translation of British and American texts on chemistry and polymers [19]. The initial stage of the system was the creation of an automated dictionary of word forms and turns of speech. This dictionary is oriented to a limited class of documents. The main criterion for selecting lexical items for the dictionary is a systematic approach (their place in the terminological system) and their frequency. The authors call the translation a word-by-word one; it has completely satisfied customers' needs for several years now.

A number of publications have also appeared on the development of a microcomputer-translator. The problems here relate more to the automated dictionary; however, one may assert that practical industrial machine translation will inevitably be connected with the automated dictionary in years to come, since the main problems of such translation are lexical ones. It is essentially in either case a matter of computer-assisted translation, the only difference being that the machine translation system assumes a larger part of man's work (at least as planned), whereas the computer plays a purely auxiliary role in the case of the automated dictionary. Among papers devoted to microtranslators, research carried out in the Minsk Institute of Foreign Languages may be noted. Extralinguistic and linguistic components are being developed. These components are the heart of the data bank of the microcomputer-translator, which translates conversational cliches from Russian to English and vice versa. 4,000 pairs of English and Russian parallelly translated Colloquial cliches have been selected. These represent frequently encountered colloquial cliches used in stereotypical city conversational situations. The data bank includes the following blocks:

(a) Russian-English and English-Russian colloquial cliche dictionaries;

(b) bilingual microdictionaries servicing individual cliches;

(c) bilingual subject field dictionaries servicing colloquial cliches for different situations in one subject field;

(d) the common bilingual dictionary servicing all subject fields, situations and cliches [20].

Such is the development of operating MT systems as described in publications from 1980 to the present.

The informational aspect of machine translation is becoming quite apparent after several years of experimental and industrial operation of the MT system. It is closely interconnected with the MT's economical specifications. The economic efficiency of this new kind of informational product becomes obvious only when there are large bulks of sufficiently homogenous texts having a specific form at the input of the system, and the number of translated texts is great. The economic efficiency depends on the quality of translation and the volume of post-editing (or inter-editing, or pre-editing). The experience gained by the All-Union Centre for Translation shows that various texts requiring different degrees of editing pass through the system. Some customers are satisfied with a rough, practically unedited machine product. In other cases, rather extensive post-editing is required, which practically nullifies the advantages of MT, putting it on a level with manual translation [21]. The All-Union Centre for Translation has a certain contingent of free-lance editors who edit the machine product.

The differentiated use of MT is economically justified only in an integrated scientific and technical information system where some forms of service can be organised on its basis, e. g. current-awareness information dissemination. However, it is clear that the machine product will not do for many traditional kinds of service. As an example, one may take the abstracting activity. There are special services which carry out this activity, and the replacement of a traditional abstract by the pilot machine translation will hardly be a smooth one for either the customer or the abstracting service. These and other questions of including MT in the traditional network of information services for scientific and technical development support need to be solved and are being worked on at present.

The All-Union Centre for Translation, within its technological scheme of translation processing and as head organisation in the scientific and technical field in the USSR and in the INTERINFORMPEREVOD international information service, is now creating a stock of machine translations; both edited and those ordered unedited by the customer. This is a special stock. Its use is less expensive. In contrast to the usual translation stock, where translations are stored in the form of hard copies, this stock is stored on magnetic media. Materials from this stock can be delivered in any form on orders of the customers. The All-Union Centre for Translation has already acquired experience in using this stock for informational servicing of customers.

Summarising, one may say that as a result of MT's development being aimed at language barriers elimination, its scientific foundations are being re-evaluated. They are being continuously enriched by MT's own development and by the contribution of related sciences, primarily, of linguistics [22]. New operating MT systems are being introduced, microcomputer-translators are coming out, automated dictionaries are being put into practice, and in-depth studies of scientific-and-engineering speech style are being conducted [23]. We are now at the stage where practice can give MT the greatest stimulus.

REFERENCES

1. Shalyapina, Z. M. Automatic translation as a model of the human translation activity. *International Forum on Information and Documentation,* 1980, 5, No. 2, 18—23.
2. Shiryaev, A. F. On specific psycholinguistic features of a functional system of simultaneous translation. In: *Translation as a linguistic problem. A collection of papers.* (Ed. by Garbovskij, N. K. et al.). Moscow: MGU Publishers, 1982, 120 p.
3. Marchuk, Yu. N. The flow-diagram of actions by a translator. In: *Text levels and methods of its linguistic analysis.* Moscow: VTsP, 1982, 21—31.
4. Andreev, N. D. Statistical and combinatorial methods in theoretical and applied linguistics. Leningrad: Nauka Publishers, 1967, 248 p.
5. Nelyubin, L. L. *Translation and applied linguistics.* Moscow: Vysshaya Shkola Publishers, 1983, 207 p.
6. Serdobintsev, N. Ya. Semantic word structure and its connotations. In: *Theory of word and dictionary entry functioning.* Saratov, 1981, 4—25,
7. Budagov, R. A. *Man and his language.* Moscow: MGU Publishers, 1974, 26! p.
8. Tyulnina, V. P. Studies into interaction between terms and commonly used words as a linguistic and translation problem. In: *Summaries of papers presented at a local conference on 'Problems in translating scientific and technical literature and teaching foreign languages at higher education institutes of technology'.* Yaroslavl', 1983, 123—125.
9. Neroznak, V. P.; Stepanskaya, S. P. Contrastive linguistics and translation problems. *Ibid.,* 223—224.
10. Leontyeva, N. N.; Nikogosov, S. L. The FRAP system and the evaluation problem of the machine translation quality. In: *Machine translation and applied linguistics,* No. 20. Moscow: MGPIIYa Publishers, 1980, 57—78.
11. Apresyan, Yu. D. An explanatory and combinatorial dictionary in the formal language model and polysemy resolution. In: *Current problems in practical implementation of automatic translation systems.* Moscow: MGU Publishers, 1982, 108-128.

12. Dorfman, A. G.; Sergeyev, V. M. The morphogenesis and latent meaning structure of texts. In: *Problems of cybernetics. Logic of reasoning and its simulation.* Moscow: Nauka Publishers, 1983, 137—147.
13. Marchuk, Yu. N. *Problems of machine translation.* Moscow: Nauka Publishers, 1983, 233 p.
14. *The contextological dictionary for machine translation of polysemantic English words into Russian.* (Compiled by Yu. N. Marchuk). Moscow: VTsP, 1976, Part I—264 p., Part II — 256 p.
15. Marchuk, Yu. N. The contextological dictionary: Use in programmed language teaching. *Computers and Humanities* (New York), 1979, 13, No. 4, 277—281.
16. Marchuk, Yu. N.; Vlasov, A. N. Some principles of automatic translation from German into Russian. *Fremdsprachen* (Berlin), 1980. No. 2, 91—99.
17. Marcuk, Ju. N.; Tihomirov, B. D.; Scerbinin, V. I. Ein System zur maschinellen Ubersetzung aus dem Englischen ins Russische. In: *Automatische Sprachubersetzung.* Darmstadt: Wissenschaftliche Buchgesellschaft, 1982, 319—336.
18. Sokolova, E. G.; Kudryashova, I. M. A syntactical analysis component in the FRAP system. In: *Current problems in practical Implementation* of *automatic translation systems.* Moscow: MGU Publishers, 1982, Part I, 167—183.
19. Abdullaeva, P. A.; Prikhod'ko, L. P. Principles of vocabulary selection for machine translation of patents. In: *Proceedings of the Seminar on 'Statistical optimisation of language teaching and computational linguistics:* Chimkent, 1980, 38—39.
20. Zubov, A. V.; Nekhay, O. A.; Tribis, L. I. Extralinguistic and linguistic components of a data bank for a translating microcomputer. In: *Problems of internal dynamics of speech norms. A collection of papers.* Minsk, 1982, 191—198.
21. Tikhomirov, B. D. Some peculiarities of machine translation application within the system of scientific and technical information. In: *Proceedings of the Seminar on 'Statistical optimisation of language teaching and computational linguistics'.* Chimkent, 1980, 132—134.
22. Kotov, R. G.; Marchuk, Yu. N.; Nelyubin, L. L. Machine translation in the early 80s. *Voprosy yazykoznaniya,* 1983, No. 1, 31—38.
23. Piotrovskij, R. G.; Rakhubo, N. P.; Khazhinskaya, M. S. *Systemic study of the scientific text vocabulary.* Kishinev: Stiinta, 1981, 156 p.

[*All references except 1, 15 and 17 are in Russian*]