MECHANICAL TRANSLATION

A. HOOD ROBERTS and MICHAEL ZARECHNAK*

Warren Weaver is the person generally credited with exciting interest in the possibilities of mechanical translation (MT) in the United States. (Earlier proposals came from a Russian, P. P. Smirnoff-Troyanski, and independently from a Frenchman, Artsouni.) His letter on the subject to some 200 of his colleagues in 1949 may be regarded as the beginning of this endeavor. The international conference on MT at the Massachusetts Institute of Technology in 1952 was an indication of the interest that scholars had taken in the formidable intellectual challenge. And the seriousness with which the United States government began to view this matter may be inferred from the fact that from 1956 to 1965 it supported 17 institutions to the tune of almost $20,000,000. (The reader interested in the history of MT is referred to Locke and Booth 1955, and Pendergraft 1967.)

In 1964 the National Science Foundation was supporting MT and MT-related efforts at Harvard, M.I.T., Berkeley, Ohio State, Wayne State, Bunker-Ramo, Texas, Pennsylvania, and the National Bureau of Standards. The Department of Defense in fiscal year 1964 was allocating for MT research nearly 2½ million dollars. At the time of this writing — fall 1970 — only five MT projects are being supported by the federal government— German-English (Texas), Chinese-English (Berkeley), the Russian-English contract with Latsec Corporation at Wright-Patterson Air Force Base, English-Vietnamese, Logos Development Corporation, and Russian-English at Wayne State University supported by the Office of Naval Research. (Current Research and Development in Scientific Documentation #14 (1966) lists 43 projects indexed under the term 'mechanical translation'. CRDSD #15 (1969) lists only 29 projects indexed under 'machine translation'.)

While MT research was enjoying its greatest financial support, there was at the same time in government funding circles concern about the expense of the research, the continued reliance on post-editing of machine output, and growing dissatisfaction with the increasingly large amount of effort necessary to achieve small advances. In April, 1964, the Automatic Language Processing Advisory Committee of the National Academy of Sciences — National Research Council, under the

* Although this article was jointly authored, M.Z. was primarily responsible for the section on the U.S.S.R. and A.H.R. for the rest. The authors appreciate the assistance of Catherine B. Hollan and Fred Bauman of the Center for Applied Linguistics staff in the preparation and editing of sections of this article.
chairmanship of John R. Pierce held its first meeting. (Members of the committee were John R. Pierce, Bell Telephone Laboratories, Chairman; John B. Carroll, Harvard University; Eric P. Hamp, University of Chicago; David G. Hays, The RAND Corporation; Charles F. Hockett of Cornell University (who resigned before the report was completed); Anthony G. Oettinger, Harvard University; and Alan Perils, Carnegie Institute of Technology.) Funds were provided for the committee by the major government agencies which had been supporting research in MT.

The position of the ALPAC might be summarized as follows: The quality of MT was such that it had to be post-edited. (A study was done by Arthur D. Little, Inc., in which machine translation output was post-edited and the errors were classified and counted. The A. D. Little group was able to tell from this study the percentage of total corrections made in each category. Approximately 200 pages of scientific Russian made up the original. Two different editors edited one set of approximately 100 pages. The second set contained 'approximately 100 pages from seven MT articles edited by at least four different editors'.¹)

<table>
<thead>
<tr>
<th>Percentage of Total Corrections Counted²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error</td>
</tr>
<tr>
<td>Word omission</td>
</tr>
<tr>
<td>A. Articles</td>
</tr>
<tr>
<td>B. Others</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Wrong words</td>
</tr>
<tr>
<td>A. Prepositions</td>
</tr>
<tr>
<td>B. Verb tense, voice, suffix</td>
</tr>
<tr>
<td>C. Others</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Russian left in</td>
</tr>
<tr>
<td>Choice</td>
</tr>
<tr>
<td>A. Choice of two</td>
</tr>
<tr>
<td>B. Choice of two, both wrong</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Unnecessary word</td>
</tr>
<tr>
<td>Symbol</td>
</tr>
<tr>
<td>Phrase not interpreted</td>
</tr>
<tr>
<td>Word order</td>
</tr>
<tr>
<td>Total Number of Corrections: 7,573</td>
</tr>
</tbody>
</table>

² Ibid., p. G-17.
A number of evaluations of MT programs have been made. Some of the more important ones are as follows:


Owing to the poor quality, the post-editor also had to be a translator. Since post-editing took about as long to do as translation from the original language, there seemed to be no reason to utilize expensive machinery in the process. Furthermore, there was no expectation that the quality of MT might improve significantly in the foreseeable future. Back in 1958 L. S. Barkhudarov and G. V. Kolshanskii in their paper, "Experiments with machine translation" (1958), asked: 'How far can the limits of machine translation be advanced? What are the inhibiting factors? Are they due to technical considerations or are they inherent in the very nature of the problem?' By 1965, there had been enough evidence gathered to be able to provide answers to these questions. All these questions have the word 'semantics' in their answer, i.e. the limits of mechanical translations cannot be advanced until one has a theoretical basis for the handling of semantics. The major inhibiting factor is the lack of semantic theory. The question of meaning — semantics — is, of course, essential to, and inherent in, the very nature of the problem. According to V. H. Yngve, 'Work in mechanical translation has come up against a semantic barrier ... We have come face to face with the realization that we will only have adequate mechanical translation when the machine can "understand" what it is translating and this will be a very difficult task indeed ...' (1964 : 275). It is true that with an analysis extending beyond the sentence boundary the chances of correct identification of polysemantic or homographic forms are increased. How-
ever, paragraph or content analysis of a given text must also be dependent on semantic mapping built into the program. Thus this is a vicious circle. The semantic component of an operational Machine Translation program in the past consisted of simple categories such as animateness versus inanimateness, abstract versus concrete, etc. Although helpful and necessary for the solving of ambiguities on the syntactic level, these few categories do not even begin to fill the need for true semantic analysis. The degree of difficulties can best be illustrated by the two following examples lying at opposite ends of a spectrum:

1. Free radicals in an excited state.
2. I saw a man in the park with a telescope.

While the ambiguity of the first example is solved easily by tagging the field as belonging to chemistry versus the fields of political science or sociology, the ambiguity of the second example cannot really be solved. A probability of correctness is the best one can hope for.

Generally speaking then, advances in machine translation have reached a plateau and further significant improvements will probably have to await the means for overcoming the semantic barrier.

But to return to the ALPAC report: its conclusions and recommendations were valid only for the United States and other English-speaking countries. Despite the obviousness of the fact that the ALPAC's findings applied only to the United States, it was surprising how many other countries, with completely different language situations — particularly with respect to amount of the world's scientific literature produced in the languages of these countries — either cut back on support for MT or toyed with such an idea for a long time before rejecting it.

The full recommendations of the ALPAC were as follows:

The Committee recommends expenditures in two distinct areas.

The first is computational linguistics as a part of linguistics — studies of parsing, sentence generation, structure, semantics, statistics, and quantitative linguistic matters, including experiments in translation, with machine aids or without. Linguistics should be supported as science, and should not be judged by any immediate or foreseeable contribution to practical translation. It is important that proposals be evaluated by people who are competent to judge modern linguistic work, and who evaluate proposals on the basis of their scientific worth.

The second area is improvement of translation. Work should be supported on such matters as

1. practical methods for evaluation of translations;
2. means for speeding up the human translation process;
3. evaluation of quality and cost of various sources of translations;
4. investigation of the utilization of translations, to guard against production of translations that are never read;
5. study of delays in the over-all translation process, and means for eliminating them, both in journals and in individual items;
6. evaluation of the relative speed and cost of various sorts of machine-aided translation;
7. adaptation of existing mechanized editing and production processes in translation;
8. the over-all translation process; and
9. production of adequate reference works for the translator, including the adaptation of glossaries that now exist primarily for automatic dictionary look-up in machine translation.

All such studies should be aimed at increasing the speed and decreasing the cost of translations and at specifying degrees of acceptable quality.

The National Science Foundation's response to this report was to notify the holders of grants for MT research that their awards would not be renewed, and it transferred responsibility for MT research from the Office of Science Information Services to the Special Projects Program of the Division of Social Sciences. Thus ended an era.

It is not the aim of the authors to give here an exhaustive summary of all MT or MT-related research. An attempt to chronicle all research labeled as MT would be of little use to the reader since at one time it was fashionable to claim that almost any type of linguistic research was going to be important for MT, thereby increasing the chance of receiving funding. Much of this research was very good, but it would be of marginal interest to the student of MT. One project which claimed to be doing 'MT' research had as its total lexicon 25 punched cards. This perhaps was 'MT research, but it was unlikely that it would be of great significance. The following centers of MT research are selected from among many; they are in the authors' estimation only some of the more important.

UNITED STATES

International Business Machines, Inc.

The Yorktown Heights IBM Machine Translation group, in existence till 1966, produced a Russian to English translation program which was exhibited at the World's Fair in New York. The program, employing a photostore disc, was used until recently by the Air Force at Wright-Patterson Air Force Base in Ohio.

Called the Bidirectional Program, IBM's translation system did not employ any syntactic analysis and no rearrangement of the sentence's syntactic components was performed. In certain well defined cases limited amounts of rearrangement took place in short uninterrupted strings on the syntagmatic level. Essentially the program translated the text word-for-word, inserting prepositions, infinitive particles, copulas, and auxiliary verbs in cases well defined by morphological markers. It did not insert English articles. In addition, it recognized several hundred special words whose translation, form, or position depended on the grammatical category or form of their immediate constituents. Other routines such as the special treatment of irregular English verbs, Russian reflexive verbs, negative particles, noun case ambiguities, resolution of certain types of homographs, etc., depended entirely
on the analysis of the immediate environment of the item under consideration. Thus the Bidirectional Program is also known as the Limited Analysis System. The dictionary used by the program consisted of approximately 180,000 entries and it included a relatively extensive system of microglossaries. Phrases formed approximately one third of the entries.

As used at the Foreign Technology Division, Wright-Patterson Air Force Base, this system was known as the Mark II. This system has now been replaced by the SYSTRAN system of Peter Toma's LATSEC Corp., La Jolla, California. (See Operating Systems.)

**Bunker-Ramo**

The 'Bunker-Ramo' or more properly the 'Garvin' approach to MT is based upon the 'Fulcrum' approach. This method is described succinctly by Garvin in “Machine translation — fact or fancy” (1967):

The 'Fulcrum' approach is based on two fundamental principles: the concept of the fulcrum and the pass method....

Let me give a grossly oversimplified illustration of the operation of the 'Fulcrum' algorithm: imagine that the following sentence were Russian and not English.

'These various compounds of copper have been treated in the technical literature on many occasions.'

In the earlier passes of the algorithm, first the nominal phrase 'these various compounds of copper' and the two prepositional phrases 'in the technical literature' and 'on many occasions' are identified and labeled as to their potential functions within the sentence (the nominal phrase is a potential subject, the prepositional phrases are potential complements). In a later pass, the verbal phrase 'have been treated' (which in Russian consists of a single word) is identified as the potential predicate. These identifications are made on the basis of the fulcrum principle. Thus, in the nominal phrase, the algorithm first identified its fulcrum, the head noun 'compounds', and then directs its searches at the modifiers ('These various') and the nominal complement ('of copper'), the structure of which had been previously identified by the algorithm. Finally, in the same late pass in which the potential predicate is identified, the algorithm fits the different sentence components together to arrive at the structure of the sentence as a whole, again using the fulcrum principle. The algorithm 'knows' that the fulcrum of a simple sentence is the predicate and therefore seeks out the predicate immediately. It then reads the grammar codes of the predicate to determine which are likely subjects and complements and, armed with this knowledge, can hunt up the remaining sentence components.

Semantic ambiguities are resolved by context searching wherever possible: the conditions in context are sought out which are likely to determine the choice of one rather than another of the equivalents of a given Russian word. Thus, in our sample sentence, the word represented by the English word 'compound' really has two English equivalents: 'compound' or 'association'. The algorithm will decide that 'compound' is the correct equivalent, because of the complement 'of copper' that follows in the immediate context.

Note that the grammatical and other information which the algorithm needs to carry out these decisions is carried in the codes that are contained in the dictionary and are made available to the algorithm by the dictionary lookup.
The Georgetown Project, defunct since 1964, had as its objective the development of an operational machine translation system capable of translating Russian into English without editing, and usable with any effective general purpose computer. Methodology was empirical; random text translations were produced periodically and the system was improved by step-by-step elimination of deficiencies observed. Among the materials produced by the project were texts and concordances of Russian scientific materials.

The Project on Linguistic Analysis at the University of California, Berkeley, under the direction of William S-Y. Wang, has been investigating the syntactic and semantic properties of Chinese and English as a step toward automatic translation between these two languages. Research has concentrated on producing a phrase-structure grammar of Chinese and developing a set of translation rules capable of operating on the syntactic analysis assigned to input sentences by the Chinese grammar. There are lexical, syntactic, and morphological translation rules: both grammar and translation rules are written according to a fixed formalism. The grammar, although considered essentially complete, is subject to revision as it is tested. The products of the project have been: a dictionary on magnetic tapes (CHIDIC) containing both the terminal rules of the grammar and the lexical translation rules, and sizable portions of a Chinese biochemical text (on magnetic tape and on punched cards).

There are two projects at the University of Texas, Research in German-English Machine Translation on a Syntactic Level, and Research in Russian-English Machine Translation on a Syntactic-Semantic Level. Both are under the direction of W. P. Lehmann. The two projects are similar in methodology and approach, both depending on two components, the Remote File Management System (RFMS) and the Linguistics Research System (LRS). The RFMS is used to create, maintain, and select subsets of both grammars and text The LRS has the capacity to retrieve the lexical, syntactic, and semantic descriptions of any language from the RFMS data base and to convert to the compact formats required for automatic lexical, syntactic, and semantic analysis and synthesis. The system is thus designed to perform three types of translation: lexical, syntactic, and semantic. Both projects have achieved essentially complete descriptions of the noun phrase in the languages...
under investigation, with the exception of relative clause constructions. Also completed are LRS programs which compact monolingual grammars, perform analysis, and display the results. The projects have produced grammars consisting of both monolingual and interlingual data, which are available in magnetic tape form (dependent on the completion of RFMS); concordances based on the respective corpora have also been produced. In addition, the Russian Master Dictionary is available for testing.

Wayne State University

The Wayne State project (Research in Machine Translation from Russian to English) under Harry H. Josselson has two stated objectives: the development of machine aids to technical translations from Russian into English, and the performance of linguistic analysis in order to accumulate sufficient data on which programmed aids may be based. Text is subjected to computer programmed syntactic analysis routines; the results are evaluated and routines are improved accordingly. The underlying technique is the fulcrum approach described above. The dictionary includes all predicative complementation patterns for the short form modifiers as well as a nearly complete list of patterns for the regular verbs. The syntactic analysis procedure HYPERPARSE, utilizes the dictionary as input in order to define and to label certain of the complementary structures as coded in the dictionary for every sentence processed; the program also creates a matrix designed to reduce the number of interpretations assignable to grammatically ambiguous sentences. Products of the project include a mathematical glossary in machine-readable form of all items in the corpus, with all items coded for grammatical and syntactic properties with English translations.

FOREIGN (EXCLUDING U.S.S.R.)*

France

Preeminent among the MT researchers in France is and has been Bernard Vauquois, the director of the Centre d’Etudes pour la Traduction Automatique (C.E.T.A.) of the Centre National de la Recherche Scientifique (C.N.R.S.). His center near Grenoble in Saint-Martin-d’Heres has been for years now one of the world’s most important centers for research in MT. Vauquois and his staff have been working on Russian-French, German-French, and Japanese-French. Their approach is a stratificational one which utilizes logico-linguistic models which proceed in the

* For a more extensive treatment of MT in Western Europe than is possible here, the reader is referred to the chapter, “Machine translation in Western Europe” (CTL 9). The authors are W. P. Lehmann and Rolf A. Stachowitz.
analytic stage from text to meaning and in the synthetic stage from meaning to text. A 'pivot' language has been developed and a series of models allowing passage from the original text (in Russian, German, Japanese) to the pivot language, as well as a model for the generation of French syntax and morphology proceeding from this pivot language has been developed.

Canada

Canada's political problems have been a powerful impetus toward English-French translation. Starting in 1965 Kathleen H. V. Booth of the University of Saskatchewan has been working to determine the optimum configuration of a man-machine system for English-French parliamentary proceedings. Much of this research has consisted of statistical analysis of the corpus and the construction of a transformation grammar.

Work in English-French machine translation by the Mechanical Translation Group at the University of Montreal has been in progress since late 1965. Early accomplishments included a system of indexing morphological bases and flexional endings and the cataloguing of bases; this dictionary of Common English was supplemented by various specialized micro-glossaries. Since 1968 the project has concentrated on applying Van Wijngaarden syntax (W-grammars [Van Wijngaarden 1968]) to mechanical translation. A team of computer scientists and programmers have worked on solving problems in defining the metalanguage which the team of linguists have used in their translation grammars and lexicon for passage from English to French via an 'intermediate language'. The work of the project, which has recently extended its interests to include German-French and Arabic-English MT as well, can be followed through the Reports of the Mechanical Translation Project (Recherche sur la traduction automatique), which are issued twice yearly. (See Appendix C for MT samples from this program.)

Japan

The Machine Translation Project at Kyoto University (principal investigator Toshiyuki Sakai) aims at constructing a system which carries out mutual translation between English and Japanese and produces as an output synthesized speech in the translated language. For methodological purposes both English and Japanese are assumed to be phrase-structure languages, and the method is therefore based on the ordered phrase structure grammar. English sentences are analyzed from the end of the sentence to the beginning; Japanese sentences are done in reverse. It is hoped that semantic information can be used to remedy deficiencies in the phrase structure analysis. Translation of scientific documents, newspaper text, school
books, and essays have been made; problems with mistranslations or uncompleted translations have been traced to the uncertainty of parts of speech, the ambiguous structure of input sentences, and the lack of rewriting rules. Materials produced during the project include an English-Japanese word dictionary of 8,000 words on magnetic tape and about 1,000 context-free-form rewrite rules, on magnetic tape and paper tape.

Other MT research in Japan is being carried out by Katsuhiko Noda and Hirohiko Ischimura at the Electrotechnical Laboratory in Tokyo (Machine Translation of Scientific English and Japanese); achievements have included the successful translation of individual sentences.

**Germany**

Research involving machine translation of scientific documents from English to German was, until recently, being carried out by IBM Deutschland at Sindelfingen. The translation system had three components: the bilingual dictionary, a series of processing programs, and the IBM 1401 data processing system. Processing consisted of three phases: analysis, structure transfer, and synthesis. A special feature of the system was that tentative results could be printed out during any step of the procedure. An account of the system, its problems and goals, can be found in Schirmer (1969).

The LIMAS research Group in Bonn is working on the development of a machine translation system based on a 'communicational' grammar. Such a grammar is intended to serve for both encoding and decoding and is concerned solely with 'content factors', those elements which serve directly as elements of information. The emphasis of the research is principally on semantic translation; descriptions of the two languages (German and English) therefore include semantic as well as morphological and syntactic information. Semantic features assigned to lexical items and syntactic structures are used to derive 'factor formulae'. The work of the LIMAS group rests on the assumption that any sentence of a language can be represented by a factor formula in the metalanguage for that language, which can then be translated into a corresponding factor formula for the target language. The LIMAS system, which is still chiefly in its theoretical stage, is also designed to handle not only metaphorical language but idioms and discontinuities as well.

Work in machine translation is also being carried out by the Germanistisches Institut and the Institut für Angewandte Mathematik at the University of Saarbrücken, which has acquired the machine translation program of the Georgetown University Machine Translation Project to use for work in Russian-German translation.
Other machine translation or machine translation related projects in non-U.S. centers that can be mentioned are: the Machine Translation project under Petr Sgall at Charles University in Prague, Czechoslovakia; the project under Ferenc Papp at the University of Debrecen in Hungary (Basic Research into the Structure of Russian and Hungarian); and the project Mechanical Translation from English into Czech under Kveta Korvasova at the Research Institute for Mathematical Machines in Prague, Czechoslovakia.

OPERATIONAL SYSTEMS

There are currently four operational MT systems which can be mentioned: the system employed by the Central Scientific-Research Institute of Patent Information and Technical-Economic Studies (TsNIIPPI) in Moscow, the SYSTRAN Machine Translation System developed by LATSEC Inc., La Jolla, California, that of the Atomic Energy Commission at Oak Ridge, Tennessee, and the system employed by EURATOM in Ispra, Italy.

The system at the TsNIIPPI is used for the translation of texts from the U.S. Patent Office Official Gazette. The algorithm employed takes the most essential grammatical relationships into account. The address method of recovery of information from the dictionary by a key is used, the key being the condensed code of the word that represents the address of information about the English word. This method is also applied to terminological phrases. Allowance is made for the possibility of the appearance of key multivalence and methods of eliminating it. The body of the text is replaced by a body of information through the operation of the address-search program, and this body of information is further processed by the grammatical-analysis program, which brings about a partial grammatical concordance of the Russian equivalents of the English words. A further program for recovery of Russian equivalents according to the directions stored in the information cells replaces the body of information with a body of Russian text. The alphabetic-numeral printing program then delivers the Russian text of the patent to the printing mechanism in letter form. A sample of the translation is given, as well as a block diagram of the operation of the system of translation programs. Total time for translation and delivery for printing of a 300 word article is estimated at about 65 seconds on a Ural-4 computer. Considerable post-editing is, however, necessary, thus reducing the economic effectiveness of the operation as a whole.

The SYSTRAN machine translation system has been used to translate scientific documents from Russian to English. In comparison with earlier systems SYSTRAN requires more time for input preparation but this is compensated for by shorter time requirements for post-editing and recomposition operations. The following general description of the system is taken from Toma, Kozlik, and Perwin (1970).
The first step consists of the preparation of the Russian text for computer translation. This is accomplished by having the text typed on an IBM MT/ST which records this information on magnetic tapes enclosed in small cartridges. To provide print format control of the English translation, special control characters are typed interspersed with the text. The text on the MT/ST cartridges is converted to IBM 360 computer processing magnetic tape utilizing the DATA Corporation MT/ST converter.

Having passed through this preparatory stage, the Russian text is ready for translation processing on the IBM 360/65. The present requirements of SYSTRAN on the 360/65 are 256K bytes of core memory, one input and one output tape and four 2314 disk storage drives for dictionary, program, and data manipulation usage.

The output from the 360/65 System is a magnetic tape with the English translation. This tape is processed on an IBM 1401 computer for the purpose of preparing the English translation listing. The 1401 computer has a special upper and lower case print chain for capitalization indication, as well as additional special characters to compose a 120 character print chain. The translation process on the IBM 360/65 comprises six computer steps. The first five steps are used to prepare the input text together with information from the Translation Disk Dictionaries for the last step which performs the actual translation.

The SYSTRAN system is modular, open ended, and capable of accepting virtually unlimited updatings without interfering with the present operational capability.

Two of the operational systems, that at the Atomic Energy Commission Laboratory at Oak Ridge and that at EURATOM in Ispra, Italy, make use of unedited MT. Both these organizations use the old Georgetown SLC program. The main use in Oak Ridge is as an alternate — poorer but faster—method of translation. The important thing in the way it is used is the requirement that a machine translation must be specifically requested by the user, whose budget is subsequently charged for the translation. Thus this system prevents the production of quantities of unwanted MT. It should be mentioned that the number of users of unedited MT at Oak Ridge is relatively small and some have a slight familiarity with Russian. This, of course, enables them to read the untranslated words which appear only in a transliterated form. (For samples of the output of the Oak Ridge system, see Appendix A.)

MACHINE-AIDED TRANSLATION

By machine-aided translation is generally meant the development of computational aids to the translator from the inception of a project and not the use of human post-editors to make up for the deficiencies in what was originally designed to have been fully automatic MT.

ALPAC's call for more research in this area has largely gone unheeded, and the research now going on is essentially a continuation of that which was being done several years ago by the Federal (German) Army Forces Translation Agency in Mannheim and the Translation Bureau of the European Coal and Steel Community.
in Luxembourg. Both these organizations use computers to produce glossaries to aid the human translator in working on individual technical documents. Both operations are generally alike, in that the translator first goes through the text, underlining those words for which translation equivalents are needed. In the Mannheim operation these underlined words are then keypunched and the keypunched information put into the computer; the output is one or more text-related glossaries. In the Luxembourg operation, on the other hand, the entire sentence containing the underlined words is keypunched and fed into the computer; the output is the sentence or sentences most nearly equivalent (in lexical items) to the input sentences, the translator thus receiving the desired items printed out with their context and in the order in which they occur in the original. The use of these computer aids has helped to increase translation productivity up to 50%.

Several research and experimental projects directed toward the development of computer aids to translators are currently being carried out in the United States. Among these the following can be mentioned:

*Technical Operations, Inc.*

A program to research the utility of machine-aided Chinese to English translation was initiated at the Systems Sciences Division of Technical Operations, Inc., in February 1968. The first phase of the program was dedicated to a study of human translations and translation aids in Chinese-English translations. The analysis of operational aspects of human translations demonstrated a great dependence on manual aids and indicated that a significant increase in speed of translation could be achieved by using machine aids. Phase 2 of the program, reported on in June 1969, concentrated on research and design of alphanumeric computer functions for aids to human translation. This phase provided not only an extensive state of the art survey of the field but an analysis of relevant computer components and a selection of appropriate elements of translation functions leading to the design and conduct of tests to evaluate these functions as well. The third phase researched language, translation, and computer operations to arrive at an experimental evaluation of the usefulness of computer-aided Chinese to English translation systems. The following functions were investigated: on-line and batch query, alphanumeric and graphic input modes, dictionary data needs in update and edit operations, computer-utilization audit, and computer-assisted language and dictionary analysis. These experiments revealed that the machine-aided translation processes tested can be up to 32% faster than the manual process; machine dictionary lookups were found to be 45% more successful than manual ones. The quality of the two types of translation was discovered to be the same. Extensive information on this third (MATE) phase can be found in Mathias and Fender (1970).
The project at the American Mathematical Society in Providence, R.I., is working on developing machine aids for the editor of mathematics publications which human translators have translated from Chinese to English. Work has concentrated on developing a system for comparing the translation of a Chinese word or phrase in the translation under consideration with the same word or phrase in past translations. The work is divided into two parts: input and output of Chinese characters into the machine in their natural form; and computer manipulations of these stored characters, for such purposes as making a concordance which can be easily consulted by the editor. The basic corpus consisted of the Chinese text of mathematics articles already published in translation by the American Mathematical Society; it is encoded in the standard four-digit telegraph code used for Chinese characters. In addition the 10,000 characters of the corpus are also stored in the graphic form at the Harvard University computer. The method employed to solve the problems of storing Chinese characters in a computer and other aspects of the project are discussed in Research on Machine Aids to an Editor of Scientific Translations, Report No. 1 (Providence, R.I., American Mathematical Society, February 1968).

IBM

Experimental work on computer-aided translation in a time-sharing environment has been carried out by the Linguistics Group at the IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y. As described by E. O. Lippmann and W. J. Plath (1970) a time sharing system has been found superior to batch processing in two respects:

1. The computer operations can be smoothly integrated with the translation process. Instead of having to wait for a batch run to be prepared, submitted and completed, the user is able to call on the resources of the computer as needed in the process of translation and get immediate responses. Rapid iteration toward the desired goal (i.e. a finished translation) can be achieved by switching back and forth as many times as required among human translation, direct dictionary lookup, editing and printing via terminals.

2. An on-line system lends itself naturally to provision of a range of translation and manuscript creation aids much wider than dictionary look-up alone. Within such a system, revising, editing and formatting as well as dictionary lookup and updating can be carried out conversationally by the user.

Time-sharing systems are seen as having the further advantage of allowing teams of translators located at different terminals to work cooperatively on large translation tasks.

Logos Development Corporation

The Logos Development Corporation of Middletown, New York, under contract
from the Rome Air Development Command, Griffiss Air Force Base, New York, is beginning to provide machine aids to English-Vietnamese translators. Essentially, what the translator gets is a text ordered glossary. The system, programmed in FORTRAN IV with some parts in machine language, keeps approximately ½ million characters in core storage. Two dictionaries are maintained. A dictionary of about a thousand high frequency words — including phrases of up to five words — is kept in core storage. The remainder stays on disc; phrases of up to 10 words are kept here. Each sentence is looked at five times. 1) First, humans scan, the text for parts which are not to be translated, that is, 'literals', for example, labels. 2) Second, lexical phrases which are stored in canonical forms are looked at and parts of speech codes are assigned. 3) The third look attempts to determine the syntactic function of the word. 4) The program then tries for further syntactic and semantic resolution; for example, the determination as to whether or not radio is a noun, verb, or adjective. 5) Finally, the Vietnamese equivalents are generated, usually by phrases and not just word-for-word. After experience has been gained in translating the simplified technical materials, the Logos Development Corporation intends to utilize this program in an attempt to translate general technical materials. (For example printouts see Appendix B.)

RAND Corporation

The most recently initiated (July 1970) work in machine aided translation is that undertaken by the Linguistics Project of the RAND Corp, under the direction of Martin Kay. In effect, however, the work is really human-aided MT. This project, which currently has three researchers assigned to it, is unique in that it has as its aim the development of programs which utilize in the translation process the knowledge of monolingual speakers of both the source and target languages. In the actual operation of the system, the speaker of the source language is presented by the computer with questions designed to disambiguate words it has looked up and found ambiguous. The same type of operation is performed with regard to syntactic ambiguities which are encountered, although the program uses a relatively sophisticated transformational analysis of syntactic operations.

The monolingual speaker of the target language functions much like a conventional post-editor with the exception that he sees the machine output on a cathode ray tube and interacts by means of a typewriter keyboard. The major difference in this type of post-editing as compared to the conventional type is that a record is kept in the computer of the changes it was found necessary to make in the machine output.

U.S.S.R.

This review will report on MT affairs in the U.S.S.R. as revealed by their own records. The first reports of Soviet experiments in machine translation appeared
in 1956 in the Russian journal *Voprosy jazykoznaniya*. In these articles credit was given to the Georgetown University experiment of 1954 as being the stimulus for their own MT research. Since that time and until the most recent article, published in *Nauchno-technicheskaja informacija* 2/11 (Nov. 1969), an extensive number of books, symposia, and articles have been published, showing that the interest in machine translation has been very much alive. They also indicate that our colleagues in the Soviet Union are much better informed about MT work abroad than we are about theirs.

Among published sources three merit special attention: (1) I. A. Mel'chuk and R. D. Ravich's *Avtomaticheskiy perevod, 1949-1963* (1967) is an excellent manual of international research, including fourteen projects in the U.S.S.R. Twenty-eight languages are mentioned on which some work on MT had been done, and papers by over 200 Soviet authors are reviewed. The book presents a complete history of machine translation in the U.S.S.R. and other countries, discussing general principles of the mechanical translation process such as MT and morphological, syntactic, and semantic structures as well as automatic abstracting and lexical coding. Among specific problems in MT mentioned by the author are: pre- and post-editing, MT dictionaries, automatic analysis of texts, phraseology, morphological analysis, syntactic analysis, segmentation routines, semantic analysis, and intermediary languages.

(2) V. Yu. Rozentsveig (1968) gives a history and state of the art of theoretical work in the U.S.S.R. He maintains that MT research must include semantic studies if it is to achieve a better quality of output. Unfortunately, practical research, such as the work of Bel'skaja, is not adequately represented. A bibliography of over 50 Soviet authors is appended.

(3) The chapter "Machine translation" in Papp 1966 (:100-26) provides an outstanding general and technical evaluation, critical remarks being balanced by a positive evaluation of Soviet achievements in the field of MT. The 25-page bibliography is very good.

As in the United States, Soviet machine translation has gone through a series of stages. The first phase was applied machine translation, the second theoretical research, and the third a fusion of the two. The theoretical research phase was brought about not by methodological opposition to applied MT as in the U.S., but rather by serious difficulties related to access to computer and programming skills.

A recently organized All-Union MT committee claims that MT will become operationally useful in the near future. This committee emphasizes that theoreticians and practitioners have worked together and that the segregation prevailing in the United States is no longer analogous to that prevailing in the U.S.S.R.

In review, we shall start with an analysis of the last report of the All-Union MT Committee. The report is signed by three leading MT researchers — O. S. Kulagina, V. Yu. Rozentsveig, and I. A. Mel'chuk. According to the authors, the do-
ominant feeling among MT researchers in the U.S.S.R., as well as abroad, was an outgrowth of the new concept of man-machine communication. Opinion was, however, divided. Those who were for it portrayed its opponents as fearful of the intervention of computers in human affairs and as skeptical of the ability of computers to solve the complex problems suspected to exist in the automatic processing of natural languages. Opponents were further reproached with not being ready to apply cybernetic principles and mathematical methods to humanitarian problems, specifically to the analysis and synthesis of natural languages.

Researchers anxious to sell MT to the public and to those in positions to provide funds tried to emphasize that although they wanted to work seriously and continue their MT research using scientific methods, they would not be able to promise quick and accurate results. It is not surprising to discover that their sponsors were not responsive to these proposals. Interest faded, and an era of disappointment with machine translation began.

One gets the definite impression from reading Soviet reports that the negative findings of the ALPAC (Automatic Language Processing Advisory Committee, U.S.A.) report of 1966 should also be listed among the causes which slowed down MT research in the U.S.S.R. This report was reviewed and translated in Nauchno-technicheskaja informacija 2/8.25-36 (1968).

Many Soviet MT researchers argued that the authors of the ALPAC report showed lack of understanding of the problems involved, confusing the scientific, theoretical, and practical aspects of the research. They believed that ineffective MT output should be taken as the basis for further support, not as a reason for stopping MT research altogether. Theoretical and practical results, they maintained, should complement each other. From their point of view, if there was a crisis in MT, it was in the financial aspects of their research, not in the working out of a practical system of translation; either they did not have enough moral or financial support, or they did not have free access to computers.

Interestingly, in their latest report, Soviet MT researchers used the findings of the Grenoble MT group to support their own ideas on the direction which MT should take in the Soviet Union. The principles which they espouse seem to be the same as those practiced by the French group. Given proper support, they believe that they can develop a commercially useful MT output with this method within the next five years.

Basic Stages and Methodological Commitments

According to the committee's evaluation of the situation, there are three basic steps to be followed:

1. Clear separation of stages in the translation process, and in particular independence of source analysis and target synthesis.
Analysis should consist of the following stages:

1) Dictionary look-up.
2) Morphological analysis.
3) 'Rough' syntactic analysis — transition from a constituent phrase tree to a tree of dependencies among the words in the phrases. A number of mistakes which occurred at stage 2 are corrected at this point.
4) Semantic analysis — transition from the tree of dependencies to a special deep structure representing the syntactic-semantic relations among the significant words in a Russian phrase. Such deep structures appear to function essentially as an intermediary language. Through these the source language represents the final result of the analysis and the initial point in the synthesis.
5) Synthesis, one stage of which is syntactic synthesis, that is, the transition from the deep structure of the source phrase to a tree of constituents of the target phrase. Markers concerning the mutual linear arrangement of the constituents are indicated within the constituents.
6-8) Morphological synthesis and printout.

II. The standard grammar has a specific form. Thus, information concerning the language is completely separated from the text-processing algorithm. This makes it possible for the linguist to concentrate on the attainment of an adequate linguistic description without worrying about the mathematical models to be used in their descriptions. The mathematical co-workers in the team take the linguistic description for granted and try to develop an optimal algorithm for it.

III. For representation of data of a natural language, formal mathematical models which make it possible to standardize the processing operations applicable to these data are worked out. All things being equal, the analytical features stated under 2) and 3) above are similar to any tripartite MT system (for example, the University of Texas MT system). In the U.S.S.R., a similar organization is characteristic of the MT team of Leningrad State University, headed by G. S. Cetjin and N. S. Filialov.

Proper correlation among mathematical, linguistic, and programming researchers should be maintained as well as correlation among procedures applicable to a practical language description of the source and target languages, general mathematical strategy, and specific programming techniques. Each component within any programming system should be separately debugged and properly fitted into the rest of the programming operations. Some researchers believe that further improvement in MT could be achieved even without introducing semantic coding. One could do this by extending the dictionary — in particular the clustered entries — and by further specification in the algorithm for synthesis and word order, etc.

Future research in MT, according to the All-Union report, should observe the following criteria:

a) The MT experiment should be 'complex', i.e. one must develop an exhaustive algorithm and have a team which will continuously work on the improvement of
that algorithm, in particular aiming at the uncovering of any problems.

b) Preparation of the next variant of the translational system — dictionaries, grammars, algorithms — should be done on the basis of available theoretical considerations and experimental results.

c) Prospective long-range research should be directed toward a future variant of the MT system.

d) Organizational procedures should aim at close contact between experimental and theoretical results, both in the field of linguistics and in the field of cybernetics.

e) In studying a natural language, one should take into consideration the hierarchical structure of connections among the various elements which constitute the text. On one side, the deep structure of any phrase should be explored in itself; on the other side, the surface structures of a set of continuous phrases should be explored since they might indicate some problems which are not discoverable within the deep structure of the phrase, and at the same time be important for the translational process.

f) In analyzing algorithms, a balance should be maintained between the level of processing in terms of dependencies and the usable information on that level as opposed to the accuracy of the analysis, i.e. there is no sense in trying to obtain a higher analysis on a lower level. Feedback of various types should also be used.

In the field of mathematical problems connected with MT, the following aspects should be specifically considered. Quite often, connections holding between words in a phrase show similarities in their features which could be optimally described by a 'local algorithm', as developed by U. I. Zhuraveiev. In addition, there is a growing tendency to utilize the concepts of logical operations such as disjunction and conjunction, and to plot the connections between them in graphic form, circuit designs and trees, and additional features of discrete objects.

The All-Union Committee stresses the following methodological commitments as those shared by a majority of experienced MT researchers throughout the world: complete separation of analysis and synthesis; the breaking down of both synthesis and analysis into levels substantiated on the content basis — specifically, the input and output for each level; the necessity of introducing a certain degree of semantic information both in synthesis and analysis; a tripartite approach for generating and processing data; use of a mathematical formalism in regard to input and output procedures; the development of a special metalanguage for describing processing operations that would help to translate linguistic statements into the absolute machine coding system; creation of language MT dictionaries of approximately 10 to 15,000 words and 2 to 5,000 idiomatic entries with full information accompanying each entry; creation of detailed grammars for the source and target languages on all levels (morphological, syntactic, and semantic); creation of a monitoring system to help the researcher to see and evaluate not only the final product but also the intermediary stages through which input data are processed.
Centers for MT activities in the U.S.S.R.

Among the more prominent MT centers in terms of number of publications and people involved the following should be mentioned in some detail.

The First Moscow Pedagogical Institute of Foreign Languages (MGPIYa), Department of Translation, headed by V. Yu. Rozentsveig. A special MT Laboratory is attached to the Department of Translation. From the very beginning, semantic problems have attracted the attention of the younger team of researchers under the general guidance of A. K. Zholkovskii. Several papers have been published by the group in Nauchno-technicheskaja informatcija. The most active researchers of the group are L. N. Iordanskaja (1964), Yu. S. Martemjanov (1959), Yu. K. Shcheglov, V. V. Ivanov (1963), and I.I. Revzin. In 1957, Revzin gave a series of lectures at the Institute under the title Introduction to the theory of machine translation and mathematical linguistics. The list of Revzin's publications is rather long, including such topics as formal treatment of German syntax, semantic problems and the methodology of research, relation between MT and information retrieval and other related topics. Revzin's favorite subject is connected with L. V. Shcherba's concept of the 'active' and 'passive' grammars. Revzin rather convincingly argues that the first is an analogue to generative grammar, and the second to recognition grammar. Revzin was among the first in the U.S.S.R. to appreciate the depth hypothesis of Yngve. His monograph Language models is well known.

In Leningrad, three teams, headed by N. D. Andreev, A. A. Kholodovich, and S. Ya. Fitialov respectively, have been working on MT problems. All three are mathematically oriented with special emphases on intermediary language and statistical observations of a pair of languages (Andreev), general language structures and their formalizations (Kholodovich), and computability of linguistic statements (Fitialov). The Kholodovich article on subclasses deserves special attention.

Professor A. A. Ljapunov and O. S. Kulagina initiated the MT program at the Steklov Mathematical Institute in Moscow. Ljapunov tends to stress the formal nature of linguistic structure. Kulagina became known through an article on language definition based on the theory of sets. Her 'linguistic operators' are appropriate for optimization of programming linguistic statements since repeatable subroutines may be used when needed. Her matrix representation of sentence parsing is also very promising.

The Institute of Precise Mechanics and Computational Techniques of the Academy of Sciences of the U.S.S.R. has also been active in MT. With Revzin's assistance, algorithms for several pairs of languages (German-, Chinese-, Japanese- and English-Russian) were designed and tested at this Institute. O. S. Kulagina, A. I. Martynova, and T. M. Nikolaeva cooperated 1) to work out a formal system for description of a language, 2) to design an algorithm for translation from the source language to the target language, and 3) to determine the optimal principles
for programming and coding of linguistic statements in the computer's memory. Furthermore, D. Yu. Panov, A. A. Ljapunov, and I. S. Mukhin reported on the feasibility of MT before the Plenary Session of the Academy of Sciences in 1956. A. P. Ershov's articles on programming and automatic translation were published abroad. I. K. Bel'skaja's first experiment in English-to-Russian translation was also carried out at this Institute.

Semantic Research in MT

It is the general experience of those who have tried to develop an MT system on a sentence basis that semantic coding is necessary to increase the quality of the random text translation. Semantic quality control can be increased in various ways by using semantic information indirectly, as in the efforts directed at increasing clustered entries from the source language (idioms, fixed phrases, technical terms, etc.). Yet all these indirect means, though very useful, do not secure an approximation to human translation for a very simple reason: an algorithm is blind. It has neither semantic eyes nor ears nor, of course, the brain to accept the string of signals and interpret it adequately, i.e. to understand the message. One of the basic problems facing semantic research is to find out whether or not the semantic layer consists of hierarchically included levels. If it does, can we assume by analogy that the terminal semantic base shall consist of a list of semantic distinctive features which by contrastive clustering serves to differentiate the morpheme, i.e. the meaning? Is it not possible to assume with equal probability that the semantic minimal units are of additive (not contrastive) nature? If the analogy between form level and content level should be further extended, one would then like to know whether or not there are 'free' and 'bound' semantic clusters, that is, given semantic qualifiers such as 'freely', 'by force', etc. Both the quantifiers and the qualifiers could be considered as the semantic constants whose role does not depend on the context. A list of semantic constants would seriously enhance the analysis of semantic syntagmatic units out of which the semantic message is constructed.

Other fundamental semantic questions are: the form of the metalanguage for representing the semantic information in the source and the intermediary language; the selection of one of the four components (phonemic, morphemic, syntactic, or semantic) in the natural language as 'central'; and the linearity of semantic display (written or spoken) and its non-linear hierarchical structure.

Soviet researchers have achieved some interesting results in their work in semantics. Mel'chuk and Zholkovskii, for example, have tried to develop a formal model of semantic paraphrasing. Thus, the sentence: 'John bought a book from Michael for two dollars' might be represented as a 'goal oriented activity' in which 'Michael is causing John to have a book' and 'John is causing Michael to have two
dollars' and 'John is causing himself not to have two dollars'. John's activity has Michael's activity as its goal and vice versa. The authors are aware of the illustrative nature of such an example. And indeed, if one wanted to say 'John stole a book from Michael', the above reasoning would also apply, which, of course, is undesirable. Some additional element would be needed such as the bound semantic qualifier 'freely' vs. 'unwillingly' (the first for 'buying', the second for 'stealing' as far as the paraphrasing of 'causing himself not to have' is concerned). In general terms the authors state that:

It is assumed that the over-all task of synchronic linguistics is to construct a functioning model of language as a model automatically producing for any given meaning all the texts that can serve as its expressions (the speaker's, or synthetic, grammar) and for any given text — the meaning(s) it conveys (the hearer's, or analytic, grammar). The speaker's meaning $\leftrightarrow$ text model is supposed to consist of several successive levels, or stages, on the way from a purely 'semantic' statement of meaning to its linguistic expression. (Zholkovskii and Mel'chuk 1969)

Semantic research in terms of MT carried out by Zholkovskii and Mel'chuk and his collaborators is based on the following three aspects of the semantic components:

1) The initial linguistic expression of the sense: from an abstract sense recording to the basic lexico-syntactic (deep syntactic) structures.
2) Linguistic paraphrasing: from basic lexico-syntactic structures (LSS) to all the LSS which are synonymic with it. This is secured by a special system of paraphrasing.
3) Syntactic manifestation of LSS: from LSS to all the corresponding surface structures.

*The MT Algorithms Published and Their Theoretical Significance*

One can develop an algorithm without having a particular well-defined model in mind. It is sufficient to solve a series of translational problems involved between the source and the target language. However, if the algorithm thus developed is operationally successful, it also has a theoretical significance. In that sense all the algorithms developed and published in the U.S.S.R. for MT purposes have theoretical significance which in turn might spur further development in algorithmic designs.

As in the U.S.A., there are 'ad hoc' labels (the so-called 'gray skies') and the 'ad omnia' tags (the 'blue sky' approach). Rozentsveig's article (1968) covering the development of MT in the U.S.S.R. during the last twenty years, seems to be on the side of the 'blue skies' (theoretical) approach. Reviewing Bel'skaja's MT algorithm from English into Russian, he says:
The program which served as the basis for MT experiments at the Institute of Precise Mechanics and Computational Techniques in the Academy of Sciences of the U.S.S.R. in 1955 went beyond simple word look-up. In addition to an English-Russian dictionary in the field of applied mathematics, it included routines for analyzing the English text and synthesizing the Russian text (Bel'skaja 1958). However, the MT algorithm which was arrived at experimentally imitated the translation process of a school child who usually translates word-for-word and who might occasionally use the immediate environment of the word for help in translating. However, they did not compare the correspondences between the systems of the target and source languages in translating. It is therefore significant that the synthesis of a Russian sentence was obtained depending on the English input information for each word and the position of each word within the English sentence. Synthesis routines were used after the word order in English had been transformed in accordance with the requirements for word order in Russian (Panov, Ljapunov, Mukhin, 1956). Such subordination of analysis to synthesis, that is, conscientious violations of the laws relating to translating correspondences between source and target languages would seem to be justified on grounds of practical considerations. (1968:172)

Bel'skaja's English-Russian algorithm has been recently published, and its output bears comparison with that of Georgetown's MT system or any similar system produced elsewhere (see Grenoble experiment, with output from Russian to French).

The MT research of I. K. Bel'skaja (1928-1964) is characterized by the following general principles:

1) An algorithm for an automatic linguistic analysis is feasible.

2) An algorithm for linguistic analysis should be based on a concrete study of the language data. It should be the result of linguistic research, not that of a mathematician.

3) An algorithm for linguistic analysis should be open to changes since its purpose is to describe and manipulate the living and changing natural language. It should be easy to update.

4) An algorithm for automatic linguistic analysis is needed for practical purposes, i.e. it should secure a precise and high quality translation. Once this precision and high quality have been achieved, it could compete with human translation.

5) The analysis requires much serious study, and the conclusions arrived at are of theoretical interest, too.

These observations are documented by D. Yu. Panov in his preface to Bel'skaja (1969). This book gives a complete algorithm for analysis of an English scientific text and its translation into Russian. Here are a few examples of the translation:

* Unfortunately in practice certain difficulties may arise.
  K sozhaleniju, na praktike neketorye zatrudnenija mogut vozniknut'.

* The KIWI-A has already been tested... and is being used for the study of heat
transfer problems resulting from operating graphite moderators.
KIVI-A uzhe byl ispytan ... i ispol'zovalsja dlja izuchenija problem peredachi
tepla, voznikajushchikh pri rabote s grafitovymi zamedliteljami...

* ... if no reshaping were applied...
... esli by nikakie preobrazovaniya no primenjal's...

Bel'skaja studied the word arrangement and the stylistic features of the English and
Russian versions on a contrastive basis. Compare the examples below:

* They had not wanted him on mission in the first place —
  Prezhde vsego, oni ne khoteli, chtoby on ekhal s missiej.

* The methods so far developed center around the linear equations.
  Metody, razrabotannye takim obrazom, sosredotochivajutaja vokrug linejnykh
  uravnenij.

The influence of cybernetics is felt in the work carried out at the Steklov Institute
of Mathematical Studies in Moscow under the general direction of Ljapunov,
Kulagina, Mel'chuk, and others. Machine translation is approached as if it were
part of a general effort to find out what areas of thought processes are subject to
automation. Here three MT algorithms were developed: from French-to-Russian
(Kulagina and Mel'chuk 1956 — see Problemy kibernetiki, 1960-61); from English-
to-Russian (Moloshnaja 1957, 1961), and from Hungarian-to-Russian {Mel'chuk
1958).

Imitating the human translation process from French into Russian, Kulagina
and Mel'chuk started with the empirical approach but later on, using the feedback
of the first output, they developed their own word classification for the French.
Kulagina's theoretical paper discussing her set theory of natural language was
based on the same experience. Both Papp and Rozentsveig stress the fact that the
dictionary was small (1200 words) and was designed for processing mathematical
texts. It was a split stem dictionary, idioms being listed in a special dictionary.
The process started with a dictionary lookup and substitution of grammar codes
for entries. Codes were checked for lexical ambiguities, which were resolved as
much as possible by contextual criteria. The French sentences were analyzed and
used as input for the synthesis of the Russian output. The French-to-Russian
system was tested on a Strela computer from 1956 to 1959. Some examples of that
output given by Papp (1966: 109) are:

I. Les relations que nous avons trouvées entre les racines et les coefficients d’une
équation conduisent assez naturellement à l’étude des formes symétriques.
Translation into Russian: Sootnoshenija, kotorye my nashli mezdju kornjami
The English-Russian system is praised both by Papp and Rozentsveig for its theoretical contribution to the general theory of language. The point is made that the English language, having few inflectional markers, must be dealt with in syntagmatic units when used for translation purposes. This in turn requires that the words be analyzed in terms of their syntactic functions. The latter are then synthesized in the target language. T. N. Moloshnaja, the creator of this system, developed the list of syntagmatic configurations (modifying the corresponding concepts of F. F. Fortunatov, Bloomfield, and Fries in regard to the explicitness of their presentation and detection on a linear level) which she considered as the set of expansion rules in the synthesis of Russian, and the set of reduction rules in the analysis of the English. Moloshnaja thought that her modifications of the MT adjusted parts of speech for English would be of no general linguistic significance, but Rozentsveig disagrees:

(The only point which should be emphasized is that syntagmatic statements have to be formal and constructive to be used with computers. These rules should indicate, for the given set of linguistic objects, the minimal components of these objects and the rules by which these objects are formed, and analysis of these objects into constituent elements — that is, both the synthesis and analysis of the objects being examined should be explicitly stated. Needless to say, such a description must have two properties — adequacy and completeness. The syntagmatic theories developed in Russia by Fortunatov and in America by Bloomfield and Fries did not possess these properties. Accordingly, the contribution of Moloshnaja consists of transforming these rules into formal rules. Thus, the concept of grammatical configuration for two and three elements from various word classes which are properly connected on the morphological and syntactic level, the list of English configurations and the list of their correspondences in Russian, the determination of the rules securing adequate English sentence parsing in terms of configurations, and the transfer to the corresponding Russian sentences were a definite contribution, not only on the level of syntactic analysis in MT, but also in terms of the linguistic theory in general.) (1968 :157)

This claim is not readily recognized or acknowledged by linguists abroad.

It was the preliminary experience with the Hungarian-Russian algorithm that prompted Mel'chuk to support a sentence-for-sentence translation based on the syntagmatic units detected in the source language.
In his monograph, Papp explains how the research on Hungarian influenced Mel'chuk's concept of intermediary language, the separation of analysis from synthesis, and the replacement of word-for-word techniques by sentence-for-sentence techniques. Papp stresses the fact that Hungarian, being an agglutinative language, raises operational problems similar to those in English and German, where the word order is unlike that of Russian and where Russian synthesis is based on the information accumulated during the parsing of the source language. The views expressed by Papp are collaborated by Rozentsveig and Mel'chuk's own description. The observations of the author of this section lead him to agree with Papp, however with such exceptions as stated elsewhere in this article.

The Intermediary MT Language

MT routines were at first worked out on a binary basis, i.e. for pairs of languages only. Thus, the corresponding transfer rules were of the type: if in the source language there is such and such phenomenon to handle, then its equivalent in the target language would be such and such. These transfer rules were either context free or context sensitive. When one tries to translate the source language X into the target languages Y, Z, W, then the question arises: should the research prepare the analysis from X to each target language separately, or have just one general analysis for X, and one synthesis for Y, Z, W, and have in addition a special language— an intermediary language— such that the X would be translated into that intermediary language (IL), and from IL to the particular target language, Y, Z, or W. It is obvious that the second way seems to be more economic. Also, there develops a sense of whether or not a universal language features list could be developed in this manner.

Judging by the Soviet descriptions of their efforts to develop an intermediary language, and also by the evaluation of these efforts by outside observers (Papp), one tends to agree with the following characteristics of these efforts:

Features occurring in a pair of languages to be mutually translatable could be represented in terms of a theory of sets as a) a union between the two sets, b) an intersection between the two, or c) a proper subset (through inclusion) of one in regard to the other member of the binary set

It was not found satisfactory to consider Russian as the language whose proper subset is any other language. Thus, version c) was dropped. With regard to versions a) and b), Soviet linguists seem to be split in their opinions. Mel'chuk's team, influenced by the experience gained from translating Hungarian into Russian tends to support the concept of the IL in terms of the union between a pair of two languages. Given any new language, the IL has to be changed if there is even one new feature in that language which is lacking in the previous two languages.
To some Soviet researchers this concept of IL seemed to resemble somewhat the concept of the proto-language which in a certain way could also be viewed as a union of the features occurring within a set of languages under comparison. The final result of such a comparison would differ only to the degree that the methodological commitments followed were incompatible or non-isomorphic.

The team of researchers under Andreev's guidance in Leningrad considered the IL in terms of the intersection between the languages to be mutually translatable. Since the intersection registers only the elements which are shared by the languages to be translated, they could be intentionally represented, and the language suitable for IL could be in effect independent of both the source and the target languages. Such an IL can have its own grammar and the vocabulary to be acceptable to the computer. The languages of the world are studied in terms of geographical and population sizes, and also in terms of present and possible future development in relation to the human race as a whole.

It is somewhat surprising to learn that the Hungarian language had such an influence upon Mel'chuk as far as the discovery of the syntagmatic units was concerned. These units seem to exist in any natural language, and have been described formally and quantitatively by Kholodovich. Mel'chuk's work, however, is rarely quoted by his colleagues. His 'Theory of subclasses' has been reviewed, by N. D. Andreev (1962).

CONCLUSION

This survey of published data on machine translation in the Soviet Union has shown that the achievements and perspectives of Soviet researchers in the field of MT are equal to if not more advanced than those of researchers in other countries. In terms of being informed about MT developments abroad, in the number of active researchers, in the magnitude and scope of topics under investigation, in the critical comparison of various MT systems and programming languages, in team cooperation between mathematicians, linguists, and programmers, and in the number of algorithms designed, Soviet MT research demonstrates a high degree of development. Finally, where the problem of limited access to computers has not hindered the testing of MT systems, the results achieved have been equal to those of similar experiments in the U.S.A.

PUBLICATIONS OF PARTICULAR INTEREST TO THE RESEARCHER IN MT


Serial Publications:
REFERENCES


IVANOV, V. V. 1963. Synchronic study of various layers in the language structure. Alma-Ata 4-5.


A. HOOD ROBERTS AND MICHAEL ZARECHNAK 2853

----- . 1961. Opytyne perevody s anglijskogo na russkij na mashine 'strela'. Moskva, AN SSSR. In-t nauch. informacii, Doklady na Konferencii po obra-
botke informacii, mashinnomu perevodu i avtomaticeskomu chteniju teksta.
Vyp. 9, 9 pp.

PANOV, D. Yu., A. A. LJAPUNOV, and I. S. MUKHIN. 1956. Avtomatizacija pere-
voda s odnogo jazyka na drugoj. Moskva, AN SSSR. Sessija po nauchnym
problemam avtomatizacii proizvedstva.


PENDERGRAFT, E. D. 1967. Translating languages. Automated language proc-

ROZENTSVEIG, V. Yu. 1968. Mashinnyi perevod. Teoreticheskie problemy so-
vetskogo jazykoznanija, 171-92. Moscow, AN SSSR.

SCHIRMER, BRIGITTE. 1969. Erfahrungen bei der Entwicklungen eines maschi-

TOMA, PETER P., LUDEK A. KOZLIK, and DONALD G. PERWIN. 1970. SYSTRAN
Machine Translation System (RADC-TR-70-168). Griffiss Air Force Base,
New York, Rome Air Development Center.

VAN WIJNGAARDEN, A., ed. 1968. Final draft report on the algorithmic lan-
guage Algol 68. Amsterdam, Mathematisch Centrum (December).

jazyka 'smisl ↔ tekst'. Mashinnyi perevod i prikladnaja lingvistika 11.5ff.
APPENDIX A.

Georgetown MT Program at Oak Ridge
A. HOWARD ROBERTS AND MICHAEL ZARECZNAK

APPENDIX A.

Georgetown MT Program at Oak Ridge
APPENDIX A.

Georgetown MT Program at Oak Ridge
APPENDIX A

Georgetown MT Program at Oak Ridge
APPENDIX A.

Georgetown MT Program at Oak Ridge
APPENDIX A

Georgetown MT Program at Oak Ridge

A. HOOD ROBERTS AND MICHAEL ZARECHNAK
APPENDIX A.
Georgetown MT Program at Oak Ridge
APPENDIX A.

Georgetown MT Program at Oak Ridge

2864  MECHANICAL TRANSLATION

APPENDIX A.
APPENDIX A

Georgetown MT Program at Oak Ridge
APPENDIX A.

Georgetown MT Program at Oak Ridge
tension regulator scale displays reading within tolerance.

&

T = 303928d

S do-
Thuộc điều khiển điều chỉnh khi trong vòng đủ dùng sao.

the regulator that assistant is installing is two inches long.

&

T = 1422504

S de-
Cấu độ điều khiển cần phải ráp là hai inch mới tính đủ.

the "mission control" switch is in the pilot’s compartment.
&

T = 352682d

S do-
Cài đặt nói điều khiển MISSION CONTROL nằm ở phòng điều khiển.

assistant will install fuel control air pressure sensing hose.

&

T = 2942362

S 4

Nguyệt phụ trợ là ráp ống dẫn nhiều cảm áp lục của

khiem soát chăng.

57. check that tension regulator scale displays reading within tolerance.
&

T = 2976447

S 5

57. Phải kiểm thuộc điều khiển điều chỉnh khi trong vòng đủ dùng sao.

pull down on any handle.
&

T = 299205d

S 6

Kéo bẻ cù tay lốp / cảm nào xung.
APPENDIX C.

Mechanical Translation Group, University of Montreal
(German-French Machine Translation)

ORIGINAL TEXTI

ES IST DIE EINNAHME BRAUSTELLE DER WELT, UND SIE
HATTE EIGENSCHAFTEN, DIE EINEN ARCHITEKTEN AN SEINEN
ERBRAUCH VERZEICHNEN Lassen. DAS BRAUSENGEBIET
AUF EINER FLUR 3000 MONETER HOHEN GLEITFLURE.
ARBEITSKRAEFTE UND BAUSTOFFE MUSEN UEBER
TAUSENDEN Von KILOMETERN HERANGEFUHRT WERDEN.
MONATE LANG SCHIEBT UNGEORDERTE DUNKLE STUFE
MONATE LANG HARSCHET SIG-UND-MANCHE NACHT STERNSCHEN.
ORAKEL FUGENT MIT EINER GESCHWINDIGKEIT BIS ZU 150.
KILOMETERN IN DER STUNDE UEBER DEN PLATZ, DIE
MITTELJAHRESTEMPERATUR LIEGT HEI-MINUS-45-GRAD
DIE BRAUSTELLE LIEGT AM ERDEPULL, KAMMER SIND DIE
VEREINIGTEN STAATEN VON AMERICA, MIT EINEN KOSTENUAFWAND
VON JAESCHIEN ZEITWURF 4 MILLIONEN DOLLAR STAMPKEN.
DIE AMERIKANER AM SÜDLICHSTEN PUNKT DER ERDE EINE
ARBEITSKRAEFTE UND BAUSTOFFE MUSEN UEBER
MITTELJAHRESTEMPERATUR LIEGT HEI-MINUS-45-GRAD
STORNOWAYE VON ZWanzig KILOMETREN AUSSTECHEN.
GESCHWINDIGKEIT BIS ZU 150.
SIERRAD. IN ZWEI JAHRN SOLL DIE SEHENLICHTE
PRODUIER TASTE AMERICA, WIE DIE PROFISIE DIE BEREITS
FERTIGGESTELLTE KUPPEL NENNEN, AUCH INNEN FERTIG
EINBERIECHT-SEIN. MAN WERDEN DIE-4 MITGLIEDER
DER AMERIKANISCHEN FORSCHUNGSTEAM, DIE BISLANG NOCH
METER EISMELDEN HAUSEN, DIE NEUEN KUPPEL:
UND LABORTAFEL IN DES ZWEITEN. KUPPEL (DIAMETRISCHER
METER) GEZIEHEN, DIE ZEIT GRAMMAT. SCHNEE UND
UMSTEIGER, DIE SICH MIT JEDEN WINTERSTORM AM POL
INNER HOHER-SCHICHTEN, BOMBEN-SCHIEH DIE 1950
AUSGESPRIENEN. WOHNHAUSEN DURC1 IHREN ANWACHSEN.
DREI 15 ZEITM.SCHREIBEN. DIET DAS ERSTE. AM PRODUIER-
DAGEVIE IST DIESEN. DRUCK MESSER UND LAEBER...
TRANSCODED TEXTI

*CELA EST LIEU-PLUS-SOLITAIRE* (CHANTIER LIBÉRÉ)
*HOMME* = VOITURE = AUTOMOBILE = LIEU UNI-ENS ARCHITÉCUTE
VERS SON/EN PROFESSION DES ESPÈCES EBER LASSER LIEN
TERRAIN DE CONSTRUCTION LIEGT SUR UNIEN(+ PEINER
3999 = MÉTRE-NAVIRE = CHAMP-GLACIAL = MATINS-BOUEUR
ET = AUTOMOBILE DE CONSTRUCTION = ÉTAT-ÉDENDRE-DE AU-DESSUS-DE
MILLIERS DE KILOMÈTRES DE DEVENIR, DEUX (DEUX)
MOISSON LARGUE (INTERROGUE LIEU+) SOLEIL = SEPTE
MOISSON LARGUE DERNIER ET = NOUIT + ENFERS-TATLES,
QUAGNALS PALAVER AVEC UNIEN(+ VITESSE JUSQUE A 329
KILOMÈTRES DERNIERS HORS AU-DESSUS-DE LIEU
PLACE, LIEU NOYEN (+) TEMPÉRATURE DE 9 ANNE= EU
LIEU PRÈS DE 94 Degré = LIEU = CHANTIER
MOISSON LIEU AU-DESSUS-DE POLE-SUD (+ PATRON-DE RECHERCHES
SORT LIEU UNIENS (+) ÉTATIENS DE SANTÉ-ECOLE, WAVE
UNIENS (+) DÉFENSE DE ANNUEL PLUSIEURE (+ MILLIONS
DOLLAR PIONNIER LIEU (+) AMÉRICAIN (+) LIEU AU-PLUS-SUD(OB)
POINT LIEU (+) TERRA UNIEN NOUVEAU (+) STATION-DE-RECHERCHES
DE-LENTIMENT-DE LIEU (+) TEMPS-LENT (+) PLACE, LIEU
POINT-DE-LIEU = INSTALLATION EST AINSI BIZARR
COMME LIEU RELUITANTE (+) D'ESPER-DE-NEIGE AUTOUR
UN ARGENTÉ BRILLANT (+) ØVALE (+) BATIMENT DE 15
Ô=ÉTRE-DE-UNIEN (+) SABLE-UNIEN (+) POLE-ATOMIQUE
AU-PLUS-TARD DANS DES LIEU TOUT-PLUS-SUD(E)
ÉNIGEUX (+) AMÉRIQUE (+) COMME LIEU (+) PIONNIERS
LIEU; AUPRÈS-DE-ACHERVE (+) COUPOLE-NOMMERS-ABUS
Ô=ÉTRE-DE-LENT (+) BOUTIQUE DENT (+) LIEU
Ô= MEMBRE S DE L'AMÉRICAIN (+) TEAM(S) DE-RECHERCHES
LIEU (+) SUGESTION ÉNIGUE (+) MÊME-UNIEN (+) BOUTIQUE DENT (+)
CAVERNES-DE-GLACE CRÉCHE (+) LIEU NOUVEAU (+) R+FABRIQUE
ET = NIVEAU TRÈS DANS LIEU ET = VÉTALES
COUPOLE (+) DIAMÈTRE 50 + MÊME ALLER-OCUPER LIEU
Ô+ TEMPS PRESSÉ DE-NEIGE ET + AMAS-DE-GLACE, LIEN
Ô+ SE AVEC CHARGE (+) TEMPÈRE-DOHIVER (+) LIEU (+) POLE
TOURS-DE-MAÎTRE-EN-EMPLOYER MÊME-DE-JUI (+) AUROURD'HUI
Ô+ 1956 EXÉCUTÉ (+) CAVERNES HABITÉES PAR LEUEN (+)
ACCROÎTANT (+) Ô+ PRÉSSION (+) BROYER ----- LIEU (+) ØVALE (+)
Ô+ RELATIONS PAR-CONTRE EST ÊTRE (+) PRÉSSION MEILLEUR (+)
Ô+ ET PLUS-LONG CRU
Ô+ Note: LIEU END-COUIN

PORTION OF OUTPUT IN NORMAL TYPOGRAPHY:

Cela est lieu-plus-solitaire, Chantier, l'hor de monde, et elle avait que j'habite, l'hor un archi-vue, vers son/pro fession découvrira, lieu
Terrain de construction est couche sur un/et peine 3000 mètre haut, en champ-glacial, moines et matériaux de construction et obliger d'autre
milliers de kilomètres en apporter devenir. Cing mois/au long brûle interrompu l'hor soleil, sept mois/au logeant jour et nuit templier-totale
Quagness haloier avec un archi-voie, jusqu'à 350 kilomètres, dans l'hor haute au-dessus-de l'hor place, l'hor voyante température de l'hor est couche plus-
de minus 45 degré. L'hor Chantier est couche en l'hor pôle-sud. Patron de construction sont lieu unis en etat de monde de mercah, avec un/a Bảnche de annuel
plusieurs/millions Dollar pilloner l'hor américain à l'hor au plus-sud/au point l'hor terre un/a nouveau/station de recherches de l'intérieur de l'hor
stellem/au glace. L'hor point-central, l'hor installe, aussi bizarre, comme, l'hor reluitant/pe désert de neige autour. Avec un/a Bảnche de annuel
bâtiement de 15 mètre hauteur, semblable un/a pôle-atomique. Au plus tard dans deux ans d'hor "au plus-sud" engloue de l'hor mercah", comme l'hor
vendredi d'hui 14 membres du américain/en tenu/de recherche, l'hor jusqu'à l'hor durée actuelle, ou des profondeur abîmes, caveraets-de-glace, crécher, l'hor ouvrir et en sditétaire et un/a Bảnche de laboratoire dans l'hor ayant deux étages couloirs (diamètre 30 mètre) aller-ocuper, l'hor temps pressé de neige et amas-de-glace, l'hor avec chaque/au
tempête d'hiver et l'hor pole toujours plus-haut emplir, manger déjà aujourd'hui l'hor 1956 crevassé/en cavernes-habitée par leur/en accroissant en
pression à broyer, ---- l'hor oval et "Engloue! par-contre est/au pression meilleur et plus-long cru"