[*Mechanical Translation*, vol.5, no.1, July 1958; pp. 8-15]

Research Methodology for Machine Translation

H. P. Edmundson and D. G. Hays, The RAND Corporation, Santa Monica, California

The general approach used at The RAND Corporation is that of convergence by successive refinements. The philosophy that underlies this approach is empirical. Statistical data are collected from careful translation of actual Russian text, analyzed, and used to improve the program. Text preparation, glossary development, translation, and analysis are described.

Introduction

THIS PAPER is the first of a series that describes the methods now in use at The RAND Corporation for research on machine translation (MT) of scientific Russian. The limitation to scientific text results from the importance of prompt, widespread distribution of Soviet scientific literature in the United States. The purpose of this series is to clarify the technical problems of computer application in linguistic research, to stimulate research in machine translation, and to encourage standardization of working materials. The present paper describes the general approach being followed, giving its philosophy and method.

The general approach used at The RAND Corporation for conducting research on MT is that of convergence by successive refinements. At each stage, automatic computing machinery is used for some aspects of translation, and for collecting and analyzing data about other aspects,

The philosophy that underlies this approach is empirical, in the sense that statistical data are collected from careful translations of actual Russian text, analyzed, and used to improve the MT program. Preconceptions about language are generally suppressed in this approach; no attempt is made to create a complete linguistic theory in advance. Nevertheless, cogent formalizations and previous knowledge of language are adopted whenever they seem useful.

The method is conveniently divided into four components:

1. Text Preparation. Russian scientific articles are pre-edited and punched into a deck of IBM cards.

2. Glossary Development. A second deck is punched, including a card for every different "word" in the text. Some pertinent linguistic information is added.

3. Translation. Using the glossary, an IBM 704 program produces a rough translation of the text. This translation is postedited.

4. Analysis. The postedited translation is studied in order to improve the glossary and the machine-translation program.

These four components of the research method are described in some detail in the present paper (see pp. 10 to 15 and Fig. 1). However, a complete exposition is contained in the RAND Studies in Machine Translation, nos. 3 through 9.

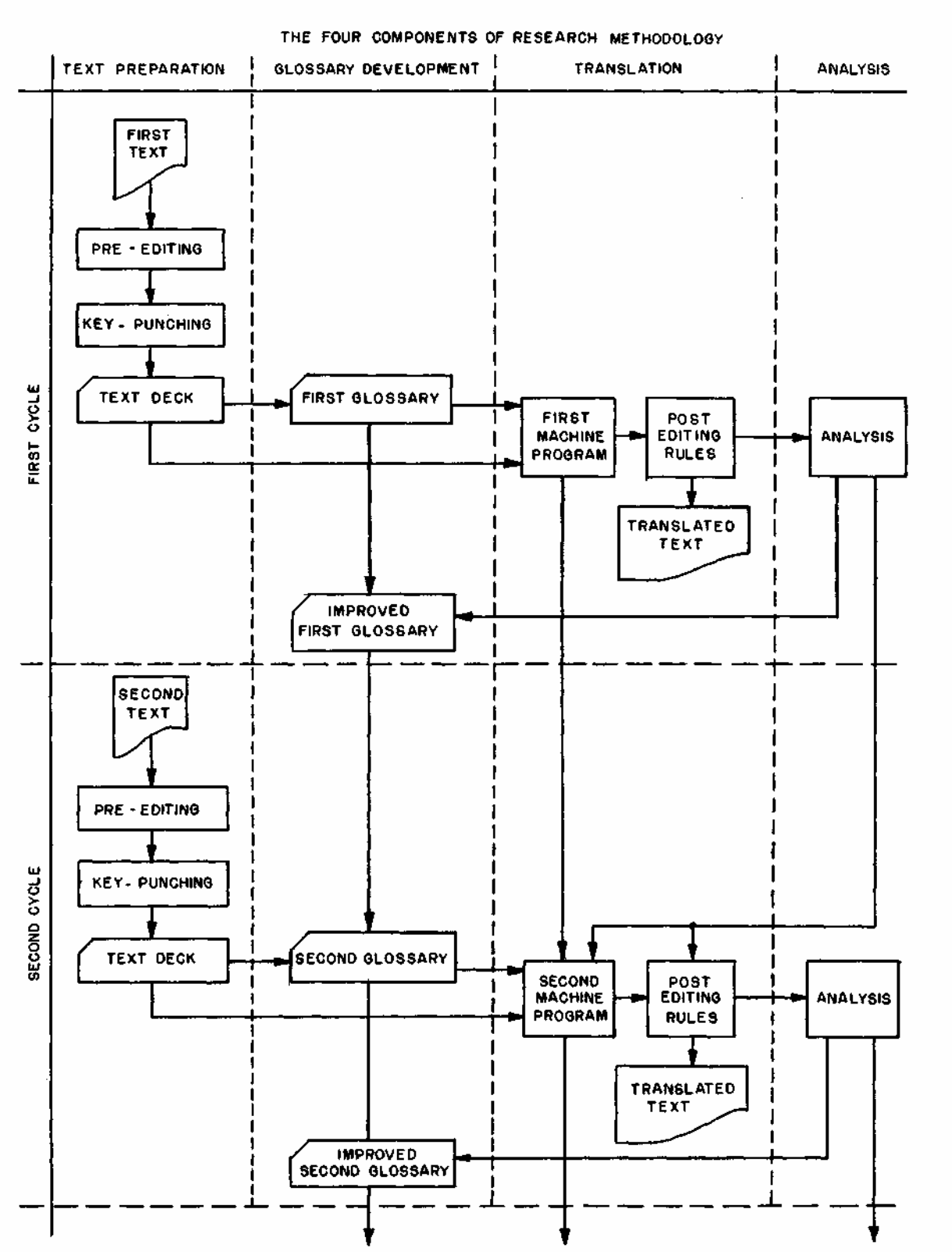
Some Definitions

It is necessary to be clear concerning the meanings of certain words that we shall use in a technical sense. This research employs a number of distinctions that are common only among linguists, and that accordingly call for special definitions.

Corpus: a group of articles or books selected for analysis.

Form: a distinctive sequence of characters. Thus every change in spelling is a change in form; "photon" and "photons" are different forms of the same word.

Occurrence (of a form): a sequence of printed characters, in a corpus, preceded and followed by either spaces or punctuation. An occurrence is identified by its ordinal position in the corpus. Hence, by definition, "photon" on page 1 and "photon" on page 2 are different occurrences of the same form.



FLOW CHART OF THE RESEARCH PROCESS

FIGURE 1

Word: a form that represents a set of forms differing only in inflection. For example, "great" and "greater" are forms of the same word, while "great* and "large" are forms of different words.

Glossary (of a corpus): a list of all the forms that occur in a corpus; grammatical and semantic information may also appear.

Dictionary (of a language): a list of all the words in the language, each represented by one form; grammatical and semantic information may also appear. A dictionary changes as the language expands and contracts.

These distinctions are necessary for precise study of language; they are used, as consistently as possible, throughout this work. Additional terms are introduced as required.

Text Preparation

The preparation of a corpus of Russian scientific text on punched cards involves selection of articles, pre-editing, design of machine codes and card formats, and keypunching.

1. Selection of Articles

The present RAND corpus consists of articles in the fields of physics and mathematics. These fields were chosen because of their importance for national security, and also because of the fact that their reputedly limited vocabularies assure a slow rate of glossary increase, which is useful in the preliminary cycles of research. Two journals are represented: Sections of the Zhurnal Eksperimental'noi i Teoreticheskoi Fiziki, which had been keypunched in a research project at the University of Michigan, furnish a valuable beginning;* in addition, articles from the Doklady Akademii Nauk SSSR are being keypunched at RAND, so that the two journals can be compared for vocabulary and sentence structure. Within the Doklady, selection is made by a scientist on the basis of substantive interest and high ratio of text to symbols and equations. A bibliography of the current RAND corpus is contained in MT Study 9.[1]

2. Pre-editing

Pre-editing is necessary for efficient keypunching; decisions are made before the keypunch operation begins, so that the operator knows exactly what to punch and in what order. The variety of characters and arrangements that is possible on a printed page cannot be reproduced on a standard keypunch machine. The pre-editor substitutes, for each nonpunchable symbol or formula, a code that can be punched. He assigns and index number to each article; to each page of the article; to each line of the page; and to each occurrence in the line. The current rules for pre-editing are contained in MT Study 4.[2]

3. Machine Codes

American punched-card machinery is not designed to process the Cyrillic alphabet; modifications are required, either in equipment or in procedure. For the present, it is most convenient to adapt procedures. Accordingly, three distinct codes for the Cyrillic alphabet are needed:

a) Keypunch Code. Special key-tops are prepared for the Cyrillic alphabet, and arranged on the keyboard of an IBM Type 026 keypunch in the pattern of a standard Russian typewriter. Each letter of the Cyrillic alphabet is punched into cards with a unique combination of holes, but these combinations are not adapted to machine sorting or listing.

b) Sort Code. The standard construction of IBM card sorting and collating machines defines a natural ordering of certain punch combinations. The RAND sort code assigns these punch combinations to the Cyrillic characters in their natural order. Thus it is possible, using standard IBM machines and standard procedures, to sort cards into Cyrillic alphabetic order.

c) List Code. The letters of the Roman alphabet, decimal digits, and a few special characters can be printed on IBM equipment. Each of these characters is printed by a unique punch combination. The RAND list code causes IBM equipment to print a Roman transliteration of the Cyrillic original. The transliteration used here was designed for convenient machine printing.

---

* Andreas Koutsoudas, the director of the Michigan project, has contributed to this RAND study as a consultant.

1. H.P. Edmundson, K.E. Harper, D.G. Hays, and A. Koutsoudas, "Studies in Machine Translation—9: Bibliography of Russian Scientific Corpus," in preparation.

2. H.P. Edmundson, D.G. Hays, E.K.Renner, and R.I.Sutton, "Studies in Machine Translation — 4: Manual for Pre-editing Russian Scientific Text," in preparation.

Of these three codes, the sort code seems most reasonable as a permanent, standard IBM code for Cyrillic characters. In the first place, the "natural" order of the punch combinations is related to the arrangement of punches in the card column, as well as to the construction of sorters and collators. Furthermore, the sort code uses one column for each Cyrillic character, whereas the list code requires as many as four columns for phonetic representations of some characters.

The keypunch code can be eliminated by mechanical alteration of the keypunch. The list code can be eliminated by construction of typewheels with Cyrillic characters for the machines used in listing. In the absence of special equipment, use of three distinct codes is unavoidable; conversions among the codes are most conveniently performed on an automatic computer.

4. Card Formats

Each occurrence of a form in the corpus, as marked by the pre-editor, is punched into an IBM card. This card contains a sequence number indicating the order of the occurrence in the corpus, punctuation marks before and after the occurrence, and the Russian form of the occurrence.

In order to record all of the information needed in translation and analysis, two cards are required for each occurrence. Both cards contain the information listed above. In addition, the first card (the translation text card) contains glossary information (see Glossary Development); the second card (the analytic text card) contains analytic information (see Translation and Analysis).

Complete descriptions of machine codes and card formats are contained in MT Study 3.[3]

Glossary Development

In accordance with the general approach of this project, the glossary is developed by increments. An initial glossary is prepared from a small corpus; examination of a new corpus leads to expansion of this glossary; and so on. Initially, the rate of growth of the glossary is large; as the process continues, the rate will decrease, but never vanish.

During each cycle, the new corpus is alphabetized on the Russian form. A summary deck is produced, containing one card for each different form; the number of occurrences of each form is recorded in this process. The new sum mary deck is mechanically matched with the old glossary, and new forms are listed for coding by linguists.

The linguist adds information to the new glossary cards as follows:

a) Grammar Code. Each form is coded for part of speech, case, number, gender, tense, person, degree, and so forth. The current RAND code has more than 1000 categories; it is described in MT Study 6.[4]

b) Word Number. Each form in the corpus is numbered automatically; it remains for the linguist to collect all inflected forms of a single word and assign a number identifying the group as a word. (See MT Study 7.)[5]

c) English Equivalents. If the new form is a form of a word in the old glossary, the English equivalents previously used are carried forward. If no form of the word has occurred before, the linguist assigns up to 3 tentative English equivalents. (See MT Study 7.)[5] His selection may be altered after postediting. (See Analysis.)

Grammar code, word number, and English equivalents are keypunched into the summary cards and then transferred to the translation text cards.

Translation

From one point of view, almost the whole research process consists of translation. In a stricter sense, however, "translation" is used to describe the two-stage process of machine translation and postediting. The process begins with the translation text deck, already containing glossary information and sorted into textual order. A 704 program produces a listing of the text as a rough translation; a postedi-tor works on this list, converting it into a smooth English version of the Russian original.

---

3. H.P.Edmundson, D.G.Hays, and R.I.Sutton, "Studies in Machine Translation—3: Resume of Machine Codes and Card Formats," August 18, 1958.

4. K. E. Harper, and D. G. Hays, "Studies in Machine Translation—6: Manual for Coding Russian Inflectional Grammar, " March 3, 1958.

5. H.P.Edmundson, K.E.Harper, D.G.Hays, "Studies in Machine Translation—7: Manual for Assigning Word Numbers and English Equivalents to Russian Forms," in preparation.

The object of this process is to produce Russian-English translations suitable for the analyses described in the following section.

1. Machine Translation

The 704 computer program for MT will eventually determine the structure of Russian sentences and construct equivalent English sentences. The program is expanded and improved as cycles of research produce more information about language, so it is impossible to give a final description of it. During the first cycle, the "machine-translation" program consisted solely of transliteration of the text and print-out of the glossary information. Analyses in the first cycle have led to the following machine routines, completed or planned:

a) Recognition of Idioms that Have Previously Occurred. An idiom is a sequence of forms that must be translated as a group, not one-by-one. This routine is ready for the second cycle.

b) Inflection of Nouns into Plural Number. The English equivalents in the glossary are generally uninflected. Hence it is necessary, when a Russian noun occurs in plural number, to inflect its English equivalent into the plural. A fairly complete routine is ready for the second cycle, but it does not take into account the fact that some forms of Russian nouns are ambiguous with respect to number. Extensions of the routine are planned to be in operation in the second cycle; these will use adjective-noun agreement to reduce the ambiguities.

c) Inflection of Verbs by Voice, Mood, Tense, Person, and Number. In English the inflection of verbs is more complicated than that of nouns. The third-person singular present tense, the past tense, the present participle, and the past participle require inflections; at times, auxiliary verbs and pronoun subjects also must be inserted. A routine to handle many inflections is planned to be in operation in the second cycle, but insertion of pronoun subjects in particular must wait for further textual analysis.

d) Insertion of Prepositions. When a Russian noun occurs in the genitive, dative, or accusative case, its English equivalent must, in most instances, be preceded by a preposition. The Russian noun may or may not be preceded by a preposition. A routine is planned to be in operation during the second cycle, which will connect Russian prepositions with their noun objects and will supply additional prepositions in English as required.

e) Selection of English Equivalents for Russian Prepositions. Russian prepositions have many alternative English equivalents. K. E. Harper, using the postedited corpus from the first cycle, has developed a classification of nouns that improves the accuracy of preposition translation. A routine is planned to be in operation during the second cycle, to select an equivalent for each preposition according to the class of the noun to which it is connected.

The computer program for machine translation has thus advanced since the first cycle began, but must be improved in every respect before machine translation is satisfactory without postediting.

The machine-translation stage concludes with the printing of a text list. The following items are printed in parallel columns:

Sequence number — Coding space —
Russian form — Grammar code —
Primary English equivalent —
Alternative English equivalents

The primary English equivalent, copied from the glossary in the first cycle, is to be modified by the machine-translation program in subsequent cycles.

The text list is designed to serve three different functions; its format economically provides for the support of these tasks:

(1) Evaluation of the Machine-translation Program. The quality of the program can be judged by reading the primary English equivalent column.

(2) Postediting. The posteditor, who must know both English grammar and the subject matter of the article can work from the English equivalents and the grammar code; he has no occasion to refer to the glossary. His notations are marked directly in the coding space; the text list then serves as a keypunch manuscript.

(3) Linguistic Analyses. The same list can be used by a linguist for structural or other analyses of the text.

2. Postediting

The posteditor inserts whatever notations are required to convert the rough machine translation into good English; his notations are analyzed in order to improve the glossary and the computer program. It is thus necessary for him to have good command of English grammar and the technical vocabulary of the scientific articles being translated. His task is to complete the work of the machine, so the rules

he follows must change from cycle to cycle as the machine-translation program develops. The following rules apply in the second cycle:

a) English Equivalents. The primary English equivalent is generally acceptable (see the following section, Glossary Refinement); if it is not, the posteditor makes one of three notations:

(1) He writes the code number of a listed alternative English equivalent in the coding space.

(2) He writes a new alternative English equivalent in the coding space.

(3) He writes a special symbol to denote that a string of occurrences is an idiom.

In one of these ways, the posteditor makes sure that the selected English equivalent is always acceptable in the context.

b) English Sentence Structure. The structure of the sentence is partially converted to English style by the machine-translation program; as that program develops in repeated cycles of research, fewer and fewer structural notes have to be made by the posteditor. Among his tasks are these:

(1) Inflection of English equivalents, or correction of the inflections made by the machine program.

(2) Insertion of English preposition codes when necessary, or correction of insertions made by the machine program.

(3) Insertion of codes giving correct English word order.

By such notations as these, the posteditor guarantees that the final product is grammatically acceptable in English.

c) Russian Sentence Structure. The posteditor indicates the connections in the sentence that make up its structure. Using such rules as the following, he writes next to each occurrence the sequence number of the occurrence on which it depends:

(1) Adjectives depend on the nouns they modify.

(2) Nouns that serve as objects of prepositions depend on the prepositions.

(3) Nouns that serve as subjects or objects of the verbs depend on the verbs.

(4) Words connected by conjunctions depend on the conjunctions.

The posteditor continues until every occurrence in the sentence, except one, is shown to depend on some other.

The selection of English equivalents and synthesis of English sentence structure was performed by the posteditor in the first cycle. Machine determination of Russian sentence structure is being initiated for the second cycle. The current rules for postediting are contained in MT Study 8.[6]

Analysis

The final component of this research methodology is analysis of the postedited translation, with the goal of refining both the glossary and the computer program. Some analyses are performed at the conclusion of each cycle; the advantages of this method include the following:

a) Compared with the preparation of a "complete" MT program before examination of any corpus, this method is more closely governed by the realities of language.

b) Compared with the translation of a very large corpus before any analysis or programming, this method is less costly, since it makes more efficient use of the posteditor's time. It is possible, by means of analyses in early cycles, to shift part of the work of corpus preparation from the editor to the computer program in subsequent cycles.

It follows that the two chief criteria for selection of analyses in each cycle are rapid reduction of the posteditor's work and selection of a corpus for each analysis large enough for statistical stability. Language problems that most often arise tend to satisfy both criteria in early cycles.

The method of analysis is empirical correlation of the posteditor's notations with the information in the glossary — word number, grammar code, and so forth. The following paragraphs describe some applications of the method.

1. Glossary Refinement

In each cycle, the glossary is enlarged by the addition of new forms and new idioms. In addition, analysis leads to improvement of the English equivalents. It is first necessary to determine, for each Russian word (i.e., set of forms) the minimal set of English equivalents required. The determination is made in the following steps:

a) A count is made of the number of occurrences for which each alternative equivalent is

6. H.P.Edmundson, K.E.Harper, D.G.Hays, "Studies in Machine Translation—8: Manual for Postediting Russian Scientific Text," in preparation.

preferred by the posteditor. The alternatives are rearranged in the glossary in order of frequency of preference.

b) In subsequent cycles, the posteditor is instructed to accept the first alternative as often as possible.

c) Secondary alternatives that are not preferred in subsequent cycles are deleted.

The English equivalents that remain are essential for accurate translation; thus it is necessary to develop criteria for choice of one of them in each context. The first task is to differentiate between the contexts in which a multiple-equivalent word is translated in different ways. The analytic text deck contains one card for every occurrence, and, alter postediting, each card is punched to show the English equivalent, and the words in the context summarized and tabulated. Presumably there are words that occur more often in the context of one preference than of the others; if such words exist, they permit differentiation of the contexts.

At least two more cycles are required before the RAND corpus will be large enough for this type of analysis. If, at that time, the data show strong differentiation of contexts, it will be necessary to construct models. One model that has been suggested is a thesaurus, or hierarchical classification of words. A model for semantic relations and a practical method for applying it are among the most important unsolved questions tions in the field of machine translation.

2. Computer-program Refinement

The general nature of the computer program is sketched in the previous section (Machine Translation). It consists of routines for determination of Russian sentence structure and construction of English sentences with equivalent structure. In early cycles, these tasks are performed by the posteditor; the purpose of analysis is to relate the actions of the posteditor to the observable characteristics of the Russian sentences, so that the computer can be programmed to take similar actions under similar circumstances.

Sentence structure is symbolized, in Russian and in English, by the following observable characteristics: word order, particles, inflections, agreements, and punctuation. For automatic computation, these characteristics are represented by word number, sequence number, grammar code, and punctuation code. Analysis consists of correlation of these characteristics of the Russian sentence with the English structural codes or structural-connection codes inserted by the posteditor.

The technique is to bring together all occurrences of form with a given grammar code — for example, all nouns in the dative plural. The analyst first tests whether any English structural code applies to all occurrences. For example, the English equivalents of Russian plural nouns must be inflected into the plural. A routine is established for English plural inflection, initiated when the Russian grammar code indicates a plural noun. Such grammatically determined routines are important, but they are few in number.

The next stage of analysis uses context of occurrence; all occurrences with a given grammar code are collected, and sorted according to grammar codes of contiguous forms. Taking the traditional rules of syntax as a guide, the analyst relates the English structural code to features of the context. The insertion of a preposition before the English equivalent of a Russian dative noun is thus related to the grammar codes of preceding occurrences. If the immediately preceding occurrence in Russian is a preposition, no additional preposition is required in English. Gradually extending the analysis over a wider context, the analyst connects dative plural nouns with preceding adjectives, preceding participial phrases, and prepositions preceding these modifiers. Syntactically determined computer routines for making the connections are written. The analyst is able to conclude that a dative noun, not connected with a preceding preposition, must be preceded by "to" in English translation. *

There are two limitations on this type of analysis. First, the structure of the sentence may be ambiguous; an adjective may be placed between two nouns with which it agrees — in Russian, it might modify either of them. It seems probable that true structural ambiguity is rare and that in most cases a sufficiently complex routine can resolve apparent ambiguities. The second limitation is that the routines are complicated by rules that are necessary for the resolution of extremely rare constructions. Since the routines must be stored in a computer of limited size, it is not practical to seek "perfect" machine translation.

---

* The example is taken from a study being conducted by D.G.Hays.

The analytic method described above is partially automatic; collection of occurrences with a given Russian grammar code, a given context, and a given English structural code is carried out by machine. With the explicit marking of structural connections planned for the second cycle, still more of the research operation becomes automatic, since it will be possible automatically to collect, for example, all dative plural nouns depending on prepositions, and to list all constructions that intervene between the preposition and the noun.

Conclusion

The RAND methodology is a system for preparing Russian scientific text on punched cards, for producing translations in analyzable form, and for exposing the relationships between the original and translated versions, semi-automatically, in such a way that translation can be programmed.

The research methodology described is, of course, designed to achieve satisfactory machine translation; the intermediate products are:

a) A descriptive grammar of the Russian language, as it is used today in scientific writing.

b) A working glossary of Scientific Russian with the English equivalents required for accurate translation.

Solutions to both conceptual and technical problems of computer application in linguistic research are given in the other papers of this series.