# GENERAL ANALYSIS TECHNIQUE

## A System of Mechanical Translation

By

Michael Zarechnak

Institute of Languages and Linguistics
Georgetown University
Washington, D. C.

## Introduction

One of three groups at the Institute of Languages and Linguistics, Georgetown University, currently engaged in a research for the mechanical translation of Russian into English, has developed under the direction of Dr.Leon Dostert, a General Analysis Technique (GAT) based on the concept of structural transfer from the source to the target language. This technique can be applied to information retrieval and translation of one language into another.

Strict linear substitution of the words of one language for those of another cannot be adopted because grammatical inter-relationships within two languages are not identical. Problems of this type as well as those of lexical (vocabulary) choice, of insertion and deletion, and of word or phrase re-arrangement are encountered in the translation of Russian into English. In the General Analysis Technique we look at the operation of translation in terms of a machine-programmable analysis and transfer of successive elements within the sentence.

There are three successive levels of linguistic analysis which are performed by the computer.[2] The first or morphemic level is concerned with the analysis of the individual word. For example, the word may take different grammatical endings. The second or syntagmatic level concerns the relations existing between immediately adjacent words. Finally, the third or syntactic level deals with locating the nucleus of both the noun and verb phrases within the sentences. These levels represent segments of the whole machine translation technique as devised by those working on the General Analysis Technique for the translation of Russian into English.*

* The present members of my group are the following: Antonina Boldyreff, Eugen Kalikin, David Korn, John Moyne, Milos Pacak, Philip Smith, and Peter Toma.

The basic feature of this method is the principle of computer-generated translation codes. The computer is provided with a series of operations allowing exhaustive analysis of the unique context. The resulting list of diacritics indicates the behavior of a word within the unique context. From this information the sentence can be translated, we have a mechanical glossary in which is only contained the inherent characteristics of the Russian word. For example, if the word is a noun, we will have codes for its gender, paradigmatic set, palatalization, semantic features, and idiom participation; and we list in the glossary only the stem of the noun.

A comprehensive presentation of the entire GAT system and its detailed routines cannot be envisaged in a short paper. I, therefore, propose to give a brief summary of a few of the routines as an introduction to the system. Those interested in further study of the system should consult, in addition to the few references listed at the end of this paper, all the Seminar Work Papers of the Georgetown Machine Translation Research project. I would, of course, be always happy to answer any enquiries concerning this system.

## GAT IDIOMATIC ROUTINE.

A type of linguistic structure which is excluded from the syntagmatic level is that of an idiomatic structure3. We define an idiomatic structure as a string of two or more Russian words which is translated into English not by its individual components but by a special equivalent structure which reflects the source concept.

In the chemical corpus under analysis we have found 271 such idiomatic structures. These 271 idiomatic structures consist of combinations of components which total 358 entries in the glossary.

The first component is separated from the rest of the components. Each of these two groups is assigned a special number on an increasing scale, and the components themselves are arranged in the Russian alphabetical order which is also used in the arrangement of the idiomatic structures.

The words forming idiomatic combinations are stored in a special idiomatic glossary. They are also stored in the main glossary carrying the idiomatic diacritic. The presence of such a diacritic will initiate an idiomatic operation such as is shown in the diagram in Figure 1. The idiomatic structure carries under each component a certain numerical code which is its identification tag. If the resulting number is one which can be produced from several source numbers, then such an ambiguous number carries a subscript which initiates an additional check to resolve the ambiguity.

When the idiomatic operation is finished, the assigned codes in English equivalents are stored in an appropriate location.

We have programmed these linguistic formulations for the translation of Russian into English and have tested the

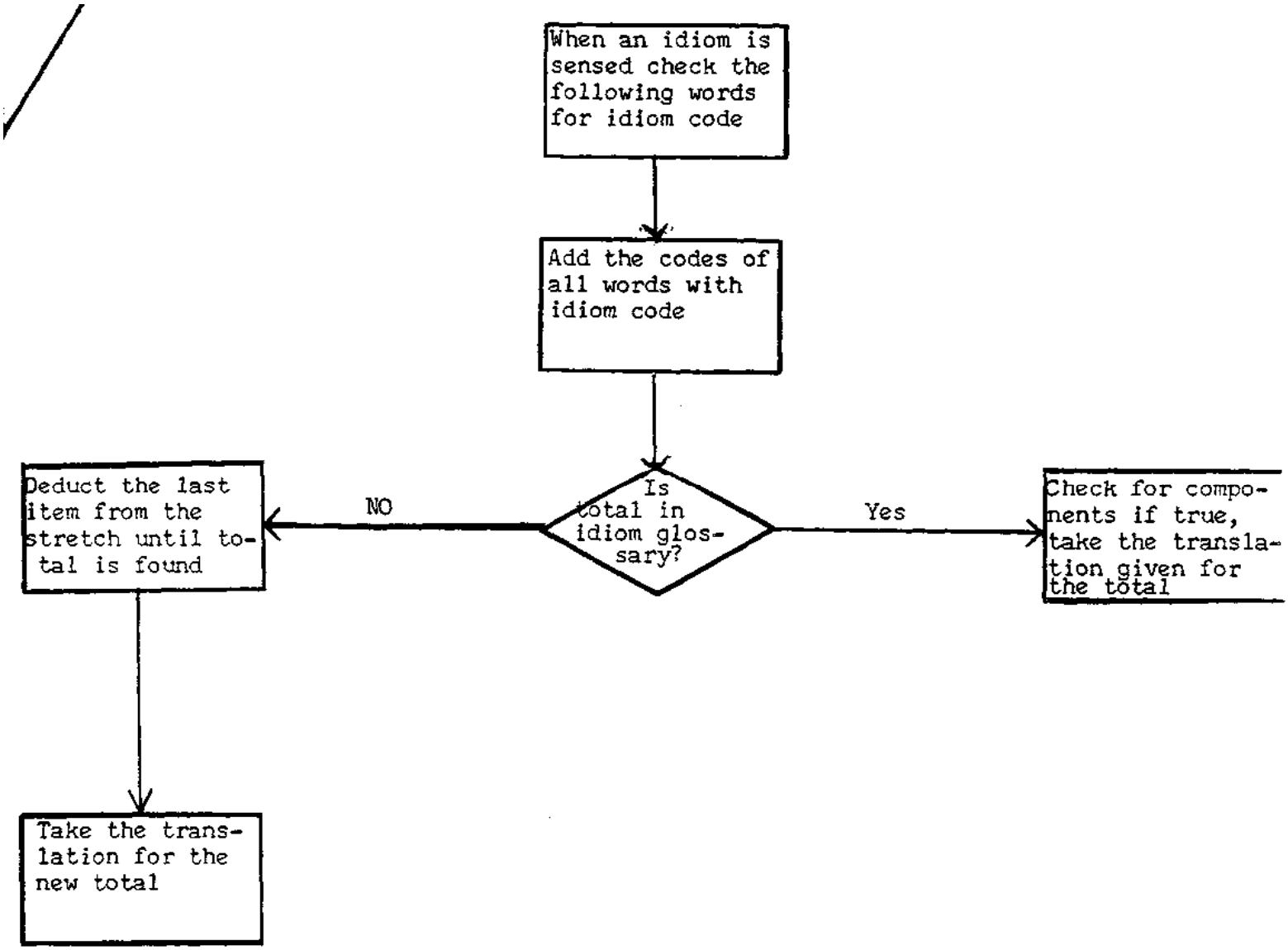



Figure 1

Major Steps in the GAT Idiomatic Routine

validity of this on an IBM 705 Computer. Since our dictionary look-up is not complicated and the addition of new words will not demand any change in the basic translation routine, we can rapidly increase the scope of machine-translated Russian scientific material. We may have to add some new operations to cover certain structural features which have not occurred in the initial corpus. But, because the formulation has been done on the basis of generalized linguistics concepts of Russian structure, we do not expect any radical changes in the existing program.

## GAT EXCLUSION ROUTINE

For the purpose of this operation, we define exclusion as a stretch of two or more words (items) within a sentence that, due to specific circumstances, can be transferred directly or translated word-for-word from the input to the output language4. An exclusion stretch is normally bound by punctuation marks. Thus members of an exclusion as well as the exclusion in its entirety are not subject to the normal morphemic, syntagmatic, and syntactic operations of the GAT technique. Examples of cases where the exclusion routine is applicable are chemical formulas, Russian words within a formula, and certain other subclauses within a sentence. .

Members of the exclusion and exclusion boundaries carry certain recognition codes. Many of these codes are automatically generated by the computer. When the computer senses these codes, it puts the exclusion routine into operation. This routine extracts the exclusion stretch and takes it to a special working area for direct transfer while the rest of the translation program continues with its normal analytic procedure.

Through extensive linguistic research, we have concluded that an exclusion must have a minimum of two items. It has

been found that any one-word stretch may be a sentence-agent and its exclusion may seriously curtail the structural analysis of the sentence.

Exclusion_Routine_Operation

Figure 2 represents the schematic flow-chart of the exclusion operation. The following steps are taken in this operation.*

1. Beginning with the first word in the sentence, check for code L in position 29.

2. When L is found, mark this item LB. This is the left boundary or the beginning of an exclusion.

3. Go to the next item, after L has been found, and check for code X in position 30. Continue this operation until all items in sequence carrying code X in position 30 have been noted.

4. Now check for code R in position 29. If there is no R in position 29 of the item immediately following the last item with an X in position 30, go back (left) to the preceding item and check for R. Continue this operation until R is found.

---

* The position referred to in the flow-chart and the following descriptions refer to positions on an 80-position IBM punch-card. For the GAT System each Russian word is key-punched on the first in a set of three IBM key-punch cards. Positions 1 to 33 are reserved for the word and the remaining positions are taken by various descriptive and analytic codes. Figure 3 shows the schematic positional distribution of a Russian word and its two English meanings adopted by the GAT. The figure represents 3 IBM punch-cards.

5. Mark the item with R in position 29RB. This is the right boundary or end of the exclusion.

6. Take all items between LB and RB for direct transfer into the target language.

7. End of routine; go to the main program.

It should be noted that the routine, as shown in the flow-chart (Fig.2), has loops to insure that every item in a given sentence is checked, all items with an X in position 30 are extracted, a stretch with less than two items is not treated as an exclusion, all items are checked until an R in position 29 is located, and finally in case of error a message is flashed without interrupting the continuous operation of the main translation program.

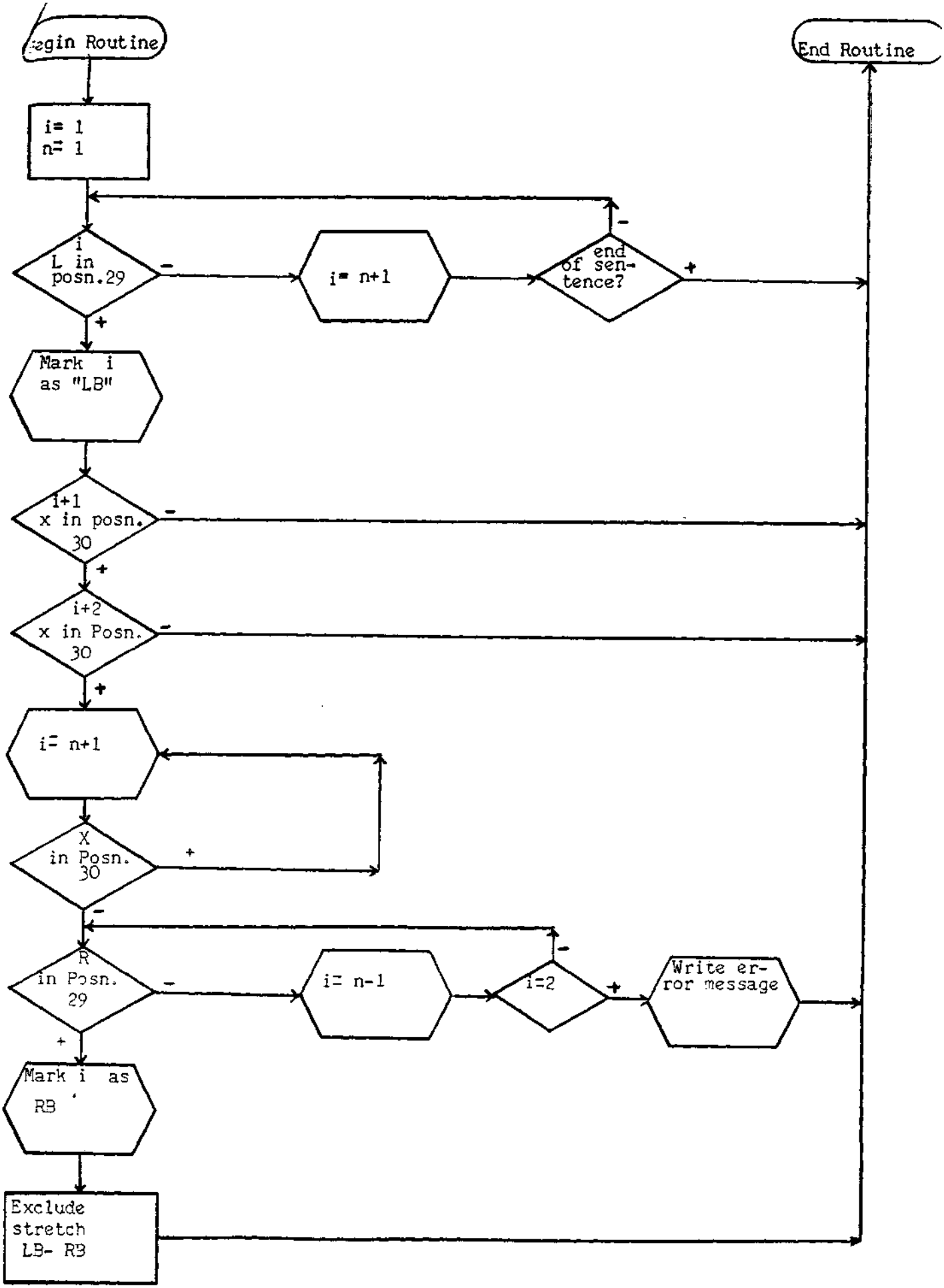



Figure 2
Exclusion Routine flow - chart

## SENTENCE SEPARATOR ROUTINE

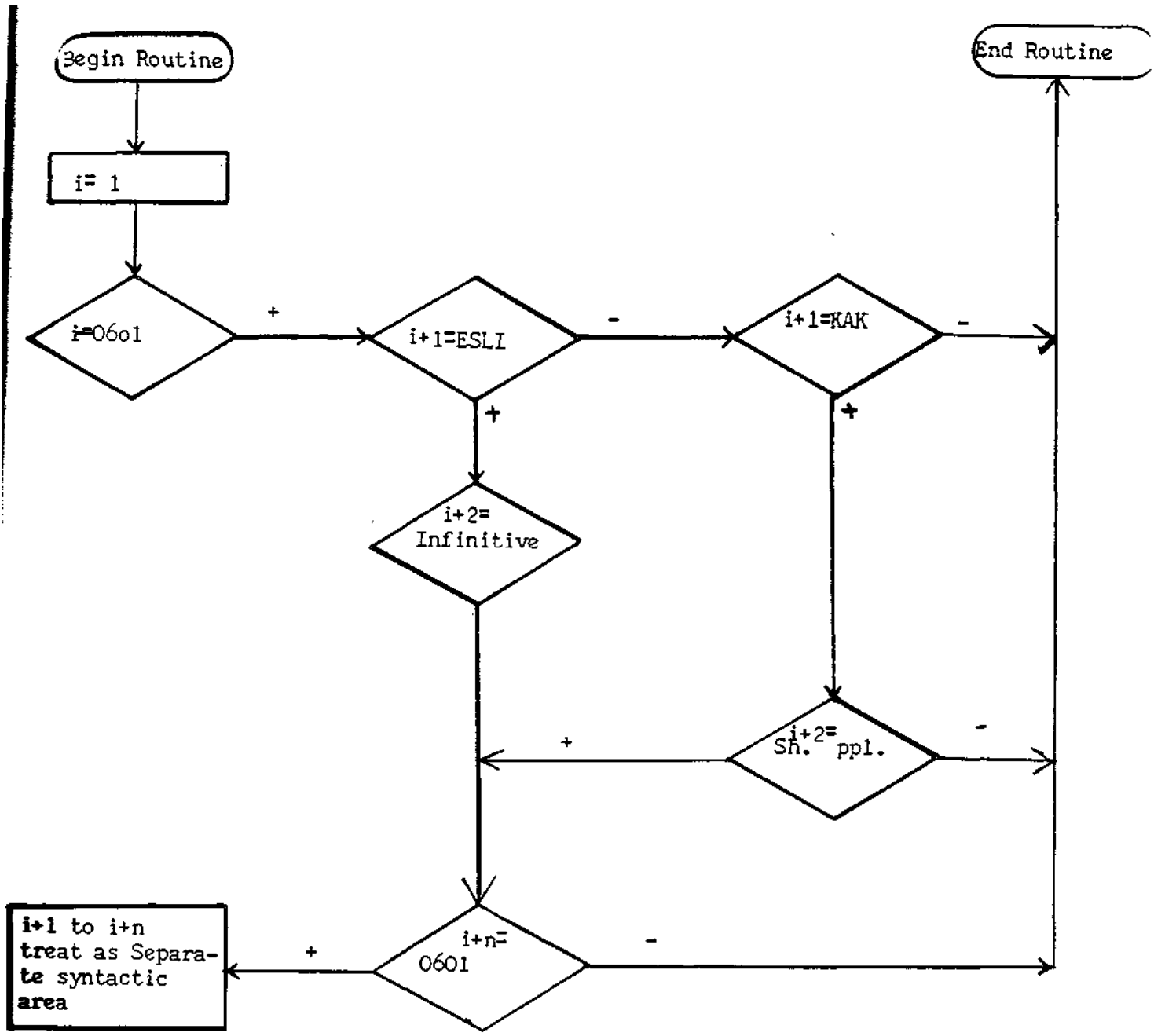
Under the exclusion routine above, I discussed a procedure for the direct transfer of an excluded stretch. In this section a technique is discussed by which a so-called parenthetical stretch is separated from the rest of the sentence and is analyzed on syntactic level independently. This exclusion area is normally recognized by punctuation marks such as period, colon, certain combination of words, etc. I have termed these in their specific above function as sentence separators which are defined as absolute sentence cuts which separate a sentence in sections, each of which is handled individually at the syntactic level. The component search for the head-word subject (H) and the predicate (P) will take place only within the defined boundaries. Figure 4 is a flow-chart for a separator operation involving a parenthetical bounded by two commas provided, that the first comma is followed by the Russian conjuctions esli (if) or kak (as). The following steps are programmed for a computer.

1. Check for comma as left boundary.
2. Check whether comma is followed by esli or kak.
3. Check whether esli is followed by an infinitive or kak is followed by a participle in short form.
4. Check whether the infinitive or short participle is followed (not necessary immediately) by a comma.
5. If the above four steps are true, treat the stretch between the two commas as a separate or independent syntactic unit.

The above five steps can be represented by the following logical notation:

$$i = 06C1 \;.\; i+l = esli \vee kak \;.\; i+2 = 21 \vee 23 \;.\; i+n = 0601 \; o \; (i+1) + (i+2) + .,.+ (i+(n-l)) \; U$$

where 0601, 21, and 23 are codes for comma, infinitive, and short participle, and U stand for the whole sentence.


Begin Routine
End Routine
i= 1
i=0601
+
i+1=ESLI
-
i+1=KAK
-
+
+
i+2=
Infinitive
+
i+2=
Sh. ppl.
-
i+n=
0601
+
-
i+1 to i+n
treat as Separate syntactic area

GAT REARRANGEMENT ROUTINE*

Rearrangement is a procedure by which the order (of words or stretches) in the input language is changed to conform with that of the target language[5]. In the GAT system recognition and instructions are keyed to the Russian sentence (see Fig.5); rearrangement is effected by moving the English items or stretches into the "working area". With each item or stretch moves its recognition code; thus the new position can be recognized on the basis of the matching codes.

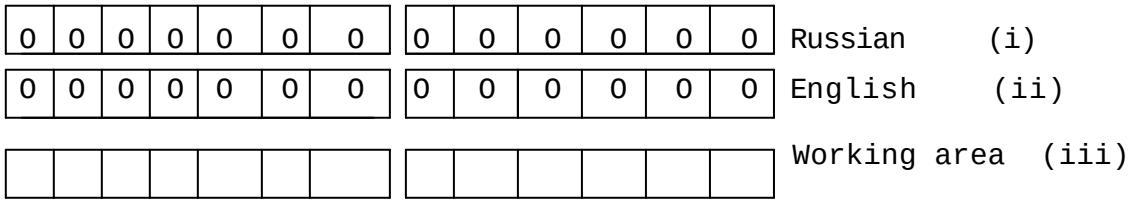



Fig. 5

The problems of rearrangement are divided into two categories of investigation: (l) linguistic research, through which it is decided under what conditions the order of words and stretches in a Russian sentence varies with the order in the English translation; and (2) a methodological study of various techniques for the functioning of this rearrangement through mechanical application. The first category determines that, for example, when a Russian phrase directly translated into English reads: "the moment beginning," it should be rearranged to read: "the beginning moment." The second category looks for a device to effect this rearrangement operation on a general basis.

We have devised a technique for the mechanical operation of the second category which is based on the conception of the

* This paper was prepared to be presented at the meeting of the American Chemical Society, at Boston, Massachusetts, April 6,1959

Boolean terms. The f(x) is expanded on 4 components with 0 and 1 coefficients. The approach is particularly convenient for work with digital computers. In examining a sentence for rearrangement purposes, first a dichotomy syntactic cut is made, that is, the sentence is divided into a subject or head-word clause (DC-H) and a predicate clause (DC-P). Then words or phrases within each clause are rearranged. There are also cases of interrearrangements between the clauses.

The following table shows some of the types of rearrangements involved in a translation process. In the chemical text we have been investigating 15 different types were discovered.

Table of Rearrangement Types

| Russian Order | English Order |
|---|---|
| a b | b a |
| a ... b | a b |
| a b c | b a c |
| a...b c | a c b |
| a . . . b c d | a c b d |

(Dots between letters indicate that the words subject to rearrangement are intercepted by other words.)

The operational procedure for this device can be represented by the following notations. Let the i in Figure 6 represent any word as the starting point in a sentence which is subject to rearrangement.


Fig. 6

The stretch between ## represents a sentence, each o represents one word, and dots (...) indicate that any number of o's can be inserted in the space.

When the rearrangement is to be effected with an immediately following or preceding item, the order is simply that of binary commutation:

i = 1000 . i+1 = 3000 o i = 3000 . i+1 = 1000 (l)

i = 2000 . i+1 = 4000 o i = 4000 . i+1 = 2000 (2)

where 1000, 2000, 3000, 4000 etc are codes for noun, verb, adjective, adverb, etc. Similar formulas have been devised for more complicated cases when the rearrangement is not between adjacent items:

(i = U4.R8).(i+n = 212xc.Vt) o (i+(n+l) = U4) .
(i+n = 212xc.Vt) (3)

When an item is governed by or governs other items, the notation — if kept on the item level (see below) — is thus:

((i) + (i+1) + (i+2) + ... + (i+n) = 5123 .
((i-n) + (i-(n+l)) + (i-(n+2)) + ... + (i-(n+p+(p+2))) =
2123 o 5123 . 2123 (4)

n and p are recognized by any mechanical device through their codes and the government Structure Procedure. Within the GAT system, however, we have adopted a different method for handling stretches. Each stretch is taken and coded as a unit. A stretch can have one or more items. On the basis of this, the notation for statement (4) could be revised to read:

j = 5123 . j-1 = 2123 . j-2 = 2123 . j-3 = 2123 o
j = 5123 . j+1 = 2123 . j+2 = 2123 (5)

where j stands for a stretch. When an item within a stretch, e.g. Russian preposition "k", is involved in the rearrangement, the notation would be:

j = "k" 5123 . j+1 = 2123 o j = 2123 . j+1 - "k"." 5123
(6)

## References

1. Zarechnak, Michael, "Report on the Work in MT at Georgetown University," Proceedings of the VIII International Congress of Linguistics, Oslo University Press, Oslo, 1958

2. Zarechnak, Michael, "Three Levels of Linguistic Analysis in Machine Translation," Journal of the Association for Computing Machinery, Vol 6, No 1, January 1959

3. Group III, "Idiomatic Structure in MT", MT Seminar Work Paper, Series B, No 3, Georgetown University, Washington, D.C., 1958

4. Pyne, Jane A.; Zarechnak, Michael, "Syntactic Transfer Procedures", MT Seminar Work Paper, No 48 and subsequent revisions, Georgetown University, D.C., May 1957

5. Moyne, John, "A System of Rearrangement Technique for Mechanical Translation, "MT Seminar Work Paper, Georgetown University, Washington, D.C., October 1958