[From: *Journal of the Association for Computing Machinery*, vol.6, no.1, 1959]

Three Levels of Linguistic Analysis in Machine Translation*

MICHAEL ZARECHNAK

Georgetown University, Washington, D. C.

Abstract. One of the research units of the current Georgetown University project in Machine Translation has developed a general analysis technique for solving MT problems. This technique is based on the concept of structural transfer from the source language to the target language. At present this research technique is being applied to Russian-English machine translation.

Various segments of the translation operation have been tested on a computer, and the results have proved helpful both to the progress of machine translation research and to the linguists formulating the technique.

The goal of the General Analysis Method, under the author’s direction, is to prove that a sentence can be handled by a machine in terms of its multi-layered constituents, so that the transfer of meaning can be adequately effected.

This paper reports the techniques, and results of several test runs on the IBM 705, of the General Analysis Technique. A three-stage program is described and linguistic explanations for these steps are provided.

The section which has been tested most extensively on the computer is the syntagmatic and syntactic analysis phase. This is concerned with establishing word groups at phrase and sentence levels.

To facilitate flow charting, linguistic statements are formulated logically. Examples of the steps leading from the linguist's concept to the computer code are discussed.

Since October 1956 a linguistic research project in Machine Translation has been in operation at the Institute of Languages and Linguistics of Georgetown University in Washington, D. C. Prior to the onset of this full-scale project, Georgetown University carried out with the International Business Machines Corporation the first practical computer test in machine translation; this experiment was conducted on an IBM701 early in 1954. At the present time Georgetown is but one of several American universities and corporations sponsoring research in the field. Research is also being done in England and the U.S.S.R.

The Director of the Georgetown project, L. E. Dostert, has consistently encouraged diversity in approach to the problem of mechanical translation. It is believed by those who have worked in the area that there is no unique solution to machine translation. Within the Georgetown project, there are currently three different groups working on Russian-to-English machine translation, and work is also being done in French-to-English machine translation. One of the Russian-to-English groups has developed a general analysis technique based on the concept of structural transfer from the source to the target language. This approach is designed to effect a complete analysis of the linguistic structure and semantic content of the Russian input text; the use of this type of analysis is not limited strictly to English translation, but has application to such uses as information retrieval and translation into other languages. It is of this General Analysis Technique (nicknamed GAT) that I will speak here.

* Presented at the meeting of the Association, June 11-13, 1958.

Although any method of translation, whether human or mechanical, requires the substitution of the words of one language for those of the other, the nature of linguistic structure precludes strict linear substitution. English words cannot be directly substituted for Russian words because the grammatical inter-relationships within the two languages are not identical. Problems of lexical (vocabulary) choice between multiple equivalents, of word or phrase rearrangement, of insertion and deletion, are some of the problems encountered when translating from Russian to English. The General Analysis Technique holds it necessary to view the translation operation in terms of a machine-programmable analysis and transfer of successively included constituents within the sentence.

The linguistic analysis can be characterized in three successive levels, or stages, which are effected internally by the computer between the input and output phases. What are these three levels? We will begin with a brief description of each, and then turn to concrete examples.

The first level concerns the analysis of the individual word. It may be inflected, meaning it may take variant grammatical endings. An example of this is given below.

The second level deals with relations existing between immediately adjacent words. The result of this analysis is a series of building blocks out of which the last level is constructed, namely the sentence. The types of building blocks for the sentence are contained within government, agreement and apposition structures.

The third level solves such problems as locating the nucleus of the noun phrase and verb phrase within the sentence. The first in most cases will be a noun in the nominative case or some substitution for it; the second takes the form of some type of verb or its substitution. This level secures enough information so that the English structural equivalent can be elicited.

In our linguistic jargon we refer to the first, second and third levels as morphemic, syntagmatic and syntactic, respectively.

These levels are not self-contained or independent stages; they represent segments of the whole machine translation technique as devised by my section of the research project. Inasmuch as language, just as any other phenomenon of the world we live in, exhibits regularity and patterning, I believe that the linguist can discover and describe the underlying concepts of this ordered system which we call language. The external expression of linguistic pattern is comparable to the time function; the irreversibility of the latter is, reflected in the importance of sequential analysis within the three levels. It is not surprising, then, that a linguist should develop the concept of a rectangular matrix to describe all the necessary operations in machine translation. (Of course I am aware of the pseudo-mathematical flavor of some of my statements, but from the linguistic point of view the matrix idea is a very practical device, exhaustive yet simple, and yielding the desired analysis.)

The rows of the matrix consist of constant operations, representing vertically for each word the operations necessary for the machine to produce all the codes to be used in translation.

The columns are shifting in character, in that the number of columns depends on the number of words in the sentence. In the Russian chemical corpus which we have used for analysis this number varies from 5 to 70 words.

The basic feature of the General Analysis method is the principle of computer-generated translation codes. Instead of the linguist supplying these in the Russian glossary, thereby having examined any possible context of a given Russian word, the computer is provided with a series of operations permitting exhaustive analysis of the unique context, and the resulting generation of diacritics indicating

Constant Locations		Content	Shifting Locations				
			Word				
Level	Row		1st	2nd	3rd	4th	Nth
A. Input	1	Russian word					
	2	Part of speech					
	3	Paradigmatic set					
	4	Gender					
	5	Idiom candidacy					
	6	English equivalent(s)					
	7	Transfer ambiguity					
	8	Case determiner					
	9	Animation					
	10	Time					
	11	Space					
	12	Voice					
B. Morphology	13	Number					
	14	Full form					
	15	Tense					
	16	Person					
	17	Case					
C. Syntagmatic	18	Interpolation					
	19	Class function					
	20	Homogeneous function					
	21	Apposition					
	22	Agreement					
	23	Noun					
	24	Verb					
	25	Prepositional } government					
	26	Adjectival					
D. Syntax	27	Exclusion					
	28	Boundary					
	29	Independent, variable					
	30	Dependent variable					
	31	Syntagmatic } rearrangement					
	32	Syntactic					
E. Output	33	English word					

FIG. 1. MATRIX Format

the behavior of a word within this unique context, the sentence being translated at the moment. We include in the mechanical glossary only the inherent characteristics of the Russian word. For example, if the word is a noun, its features will be coded in terms of its gender, palatalization, paradigmatic set, idiom participation, and semantic features. We list in the glossary only the base or stem of the noun, and thus avoid the redundancy involved in listing the noun in all its inflected forms.

Now let us turn to the details of the matrix format, in figure 1.

Section A contains the input data, taken from the Russian glossary, and located in rows 1 through 12.

Section B represents analysis level 1, the operation effecting morphemic analysis (putting grammatical suffix and stem together). The results of this operation are recorded in rows 13 through 17.

Section C is the second level of linguistic analysis. It contains the locations for storing codes pertaining to relations between immediately adjacent words on the basis of the discovered linguistic structures of agreement, government and apposition. All of these codes are generated by the computer program and stored in rows 18-26.

Section D is analysis level 3, the syntactic operation. When the subject of the sentence is located, an appropriate code is stored at this location. Furthermore, the cuts between noun phrase and verb phrase are registered here. The results of this operation are stored in rows 27-32.

Section E is the output working area, where the English equivalent is synthesized. The English stem is selected to replace the Russian word stem, and the Russian grammatical ending is replaced by an appropriate English ending or by the insertion of a preposition. The result is stored in row 33.

Any Russian word is subject to analysis at all three levels, but positive results will be recorded at only a portion of the vertical locations, depending on the nature of the given word.

We will now give concrete examples taken from sections B, C, D, and E.

The first is from section B, morphology.

Let us assume, for example, that the Russian input contains six letters. The first five of these are found in the glossary as a possible word stem. All six letters (the full form of the word) are not located in the stored glossary. The sixth letter is -E, which is found in the list of possible endings, or suffixes. At this point the ending E operation goes into effect and follows the sequence outlined in the flow chart in figure 2.

By way of explanation of the symbols used in the flow chart, the linguistic "parts of speech" are designated as follows: U-l: noun; U-2: verb; U-3: adjectival; U-4: adverbial; U-5: preposition; U-6: conjunction; U-7: particle; U-8: punctuation; U-9: non-Cyrillic forms (such as numerals, Romanized expressions).

The example to be presented from section C, the syntagmatic phase of the translation program, is a portion of the agreement operation. Agreement is one of the three linguistic structures which characterize immediately adjacent words. Let us describe briefly the nature of these three structures.

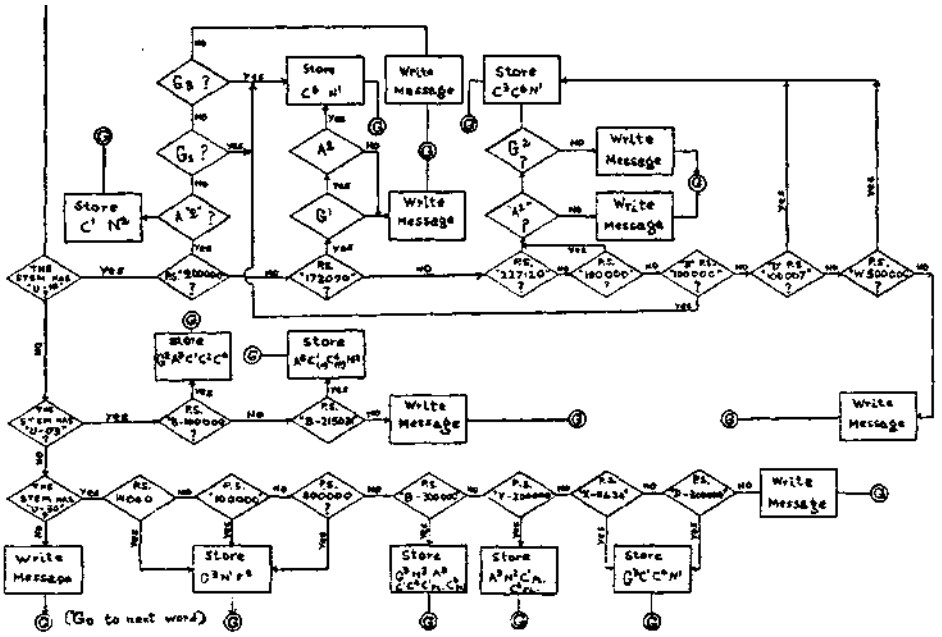

FIG. 2. E Operation

By government structure, we understand a state of predictability of the inflectional case of a second word, on the basis of the preceding case determiner in the first word. Take for example the choice between the forms "they" and "them" in the English sentence, "I saw____this morning". The native speaker will of course select the word "them". Its form is said to be governed by the preceding verb.

By apposition structure, we refer mostly to the relationship of an adverbial form to some particular word in the sentence. The two together comprise a meaningful set, yet there is no formal grammatical relationship of government or agreement to mark the bond. An adverbial item in Russian can relate to a noun, verb, adjective, or another adverb. A similar situation exists within the English language; for example, in the sentence, "I saw them briefly this morning", where the -ly form as an adverb modifies the verb.

Now let us discuss agreement structure, from which a concrete example of the mechanical operation will be given. By agreement structure, we mean an identical distribution of some grammatical feature between two words. Compare the phrases, "this young tree" versus "these young trees." The words "this" and "these" are not mutually replaceable, nor are the words "tree" and "trees", whereas the word "young" does not participate in this type of grammatical game. The common feature exhibited between "this" and "tree" and between "these" and "trees" is the concept of singular versus plural. In Russian the word "young" would also share this feature.

Many other combinations in addition to that of adjective and noun enter into

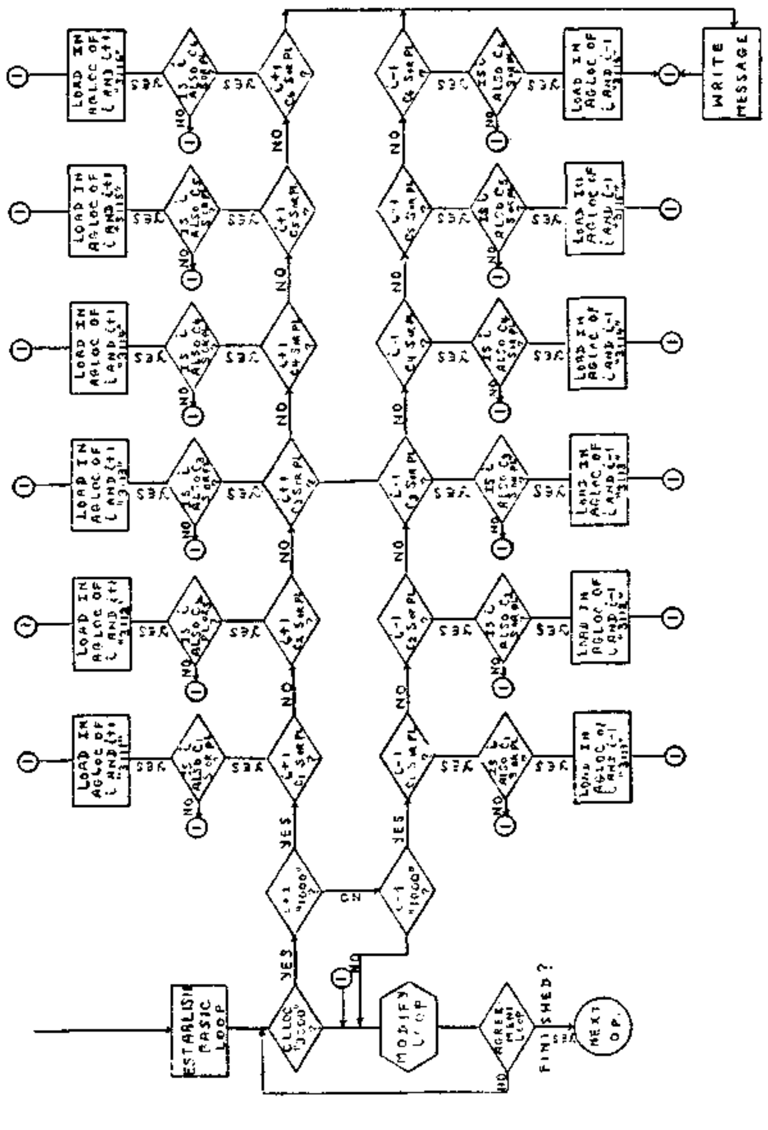

FIG. 3. Agreement Operation

agreement relationships, but we shall consider for the purposes of this discussion only the adjective plus noun set. Such pairs function as single units within the total sentence structure and need to be identified not only for purposes of grammatical translation but for operations of rearrangement for the English output.

The job of the computer program is to locate and identify agreement structures as they appear in the context of a Russian sentence. The program then attaches an appropriate diacritic to both members of the structure, and this diacritic is designed to indicate the nature of the agreement relation, the classes participating in the structure, and the grammatical features which control the relation.

The computer program proceeds as follows. It checks beginning with the first word of the sentence for the occurrence of an adjective. When a member of the adjective class is located, a check is made for a noun occurring immediately to the left or to the right. If so, the grammatical features of the adjective and noun are compared to discover whether they participate in an agreement relation. If all the necessary criteria are satisfied, a diacritic is stored at a particular address under each member of the structure. This is a four-digit diacritic. The first and second indicate the classes of the participating members. The third digit indicates the type of grammatical relationship, and the fourth records the inflectional case which characterizes the structure. For example, upon encountering the words “ximiceskix soedinenii”, meaning “chemical compounds”, the computer will store under both words the diacritic 3112; 3, 1 mean adjective plus noun, the third digit 1 means regular agreement, and the final digit 2 means the genitive case.

A flow chart for a portion of the agreement operation is given in figure 3.

The programmability of these linguistic formulations has been confirmed by

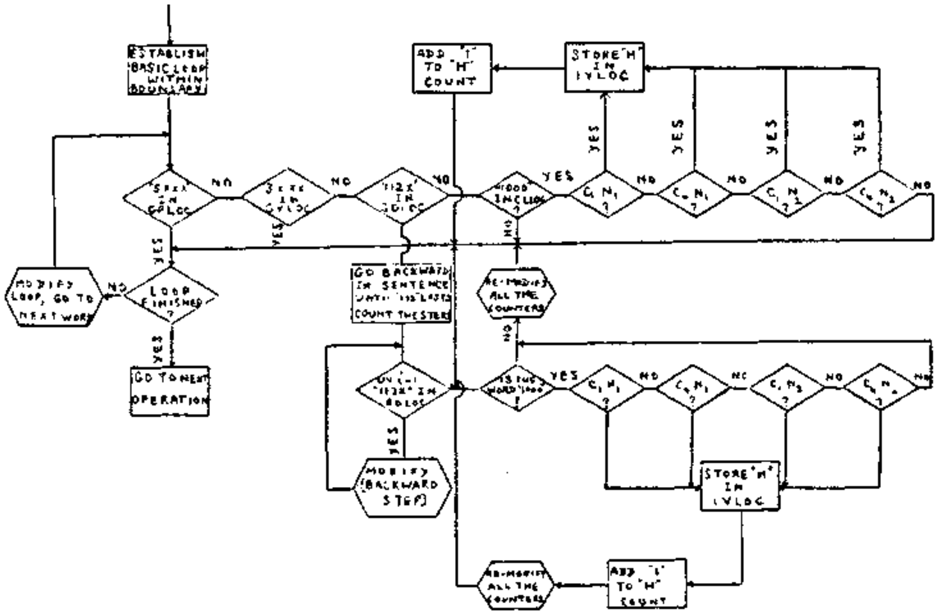

FIG. 4. Subject Operation

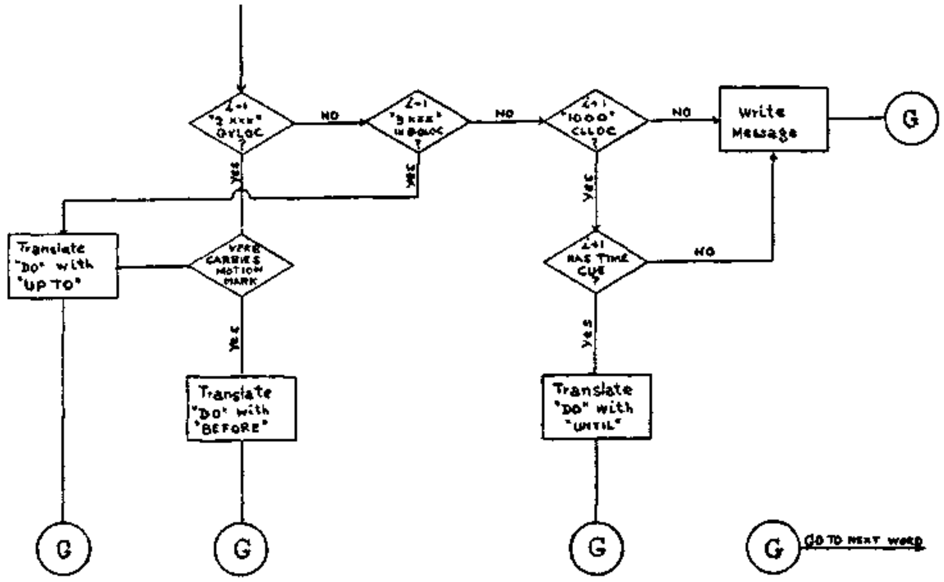

FIG. 5. DO Operation

several runs on the IBM705 computer. Tests have included the idiom glossary look-up, and detailed syntagmatic and syntactic operations from levels C and D.

One sentence which has been analyzed in the partial tests is the following:

НА ЛИНИИ ЛИКВИДУСА СИСТЕМЫ, ИССЛЕДОВАННОЙ ДО 65 МОЛ. % KCL /ДАЛЕЕ ИЗУЧЕНИЮ ПОМЕШАЛА ВЫСОКАЯ ТЕМПЕРАТУРА ПЛАВЛЕНИЯ СМЕСИ/, ИМЕЕТСЯ РЯД ВЕТВЕЙ КРИСТАЛЛИЗАЦИИ ИНКОНГРУЭНТНО ПЛАВЯЩИХСЯ ХИМИЧЕСКИХ СОЕДИНЕНИЙ.

J. General Chem., Moscow, 22 (1952).

The code generated by the computer and stored under the words of the sentence were utilized by the program to produce the following English translation:

On the liquid curve of the system, studied up to 65 mol. % KCL (the high melting point of the mixture prevented further study), there is a series of branches of crystallization of incongruently melting chemical compounds.

The codes produced under each word are as follows:

NA	5000				5126		
LINII	1000			1122	5126		
LIKVIDUSA	1000			1122	5126		
SISTEMY	1000			1122	5126		
,							
ISSLEDOVANNOI							
DO	5000				5122		
65	3000	3002	3112		5122		
MOL.	3000	3002	3112		5122		
%	3000	3002	3112		5122		
KCL	1000		3112		5122		
/							
DALEE	4000					413P	E

IZUCENIH	1000			2123	413P	E	
POMEWALA	2000			2123		E	Pr
VYSOKA4	3000	3111				E	
D							
TEMPERATURA	1000	3111	1122			E	H
PLAVLENI4	1000		1122			E	
SMESI	1000		1122			E	
/							
,							
IMEETS4	2000		1122				Pr
R4D	1000		1122				H
D							
VETVEI	1000		1122				
KRISTALLIZAQII	1000		1122				
INKONGRU3NTNO	4000				433P		
PLAV45IXS4					433P		
XIMICESKIX	3000	3112					
SOEDINENII	1000	3112					

Section D, the syntactic level, is designed for rearrangement operations within noun phrases and verb phrases as well as between the two. It is necessary for the computer to identify the head word of the noun phrase and the head word of the verb phrase. This routine makes possible a compression of any Russian sentence type into one of the following: 0-0, 0-1, 1-0, 1-1, 1-2, 2-1, 0-2, 2-0, or 2-2. The first digit of each set refers to the head word of the noun phrase, and the second digit to the head of the verb phrase. Zero means absence of the form, 1 means single occurrence, and 2 means "more than single occurrence." Thus a Russian sentence containing two subject noun phrases and one verb phrase is represented as the type 2-1. We refer to the head of the noun phrase as the independent variable, and to the head of the verb phrase as the dependent variable.

A flow chart for the operation which identifies the head word of the noun phrase is given in figure 4.

Finally, we present an example from the transfer procedure, to demonstrate how semantic criteria are used in this phase. We store with each word three semantic cues, if these are inherent in the word. Thus the preposition "DO" may be translated into English in different ways, depending on certain semantic criteria of time and space in the immediate context of the preposition.

Figure 5 is the flow diagram for the translation of the preposition "DO", indicating the method of choice between multiple equivalents in English.

In conclusion, I would like to make a few remarks concerning current planning for continued test runs on the computer. We expect to translate a continuous corpus of more than 1000 sentences before the end of the calendar year. If this translation is successful, we can rapidly increase the scope of machine-translated Russian scientific material, since our dictionary look-up is not complicated and the addition of new words will not demand any change in the basic translation routine. A greatly expanded corpus may require the addition of some new operations covering certain structural features which have not occurred in the initial corpus. Because the formulation has been done on the basis of generalized linguistic concepts of Russian structure, we do not expect any radical changes in the existing program, no matter how many sentences we put to the test.