CURRENT RESEARCH AT GEORGETOWN UNIVERSITY

M. Zarechnak and A. F. R. Brown
Georgetown University

The research in machine translation which started at Georgetown on a broader basis in the fall of 1956, after the initial three years of work, was text-focused in Russian; i. e. , the structural data and the lexical materials were derived from a selected continuous text in the field of organic chemistry of about 30, 000 words for the purpose of effecting machine translation into English. Simultaneously work was begun to translate French in the field of physics into English.

In the case of Russian, three groups worked along separate lines. Only one of the groups, now designated as the "Georgetown Automatic Translation" staff (formerly; "General Analysis Technique") carried out its work exclusively on the basis of the decision to have the research focused on a selected corpus. In the case of French-to-English, the initial analysis was based on short texts in physics, which were increased to a total of 200,000 keypunched words.

Georgetown Automatic Translation

In its present state, the Georgetown Automatic Translation for Russian-to-English consists of the following parts:

A. The dictionary
B. The algorithmic operations, which include:
   1. Dictionary lookup
   2. Morphological analysis
   3. Syntagmatic analysis
   4. Syntactic analysis
   5. The transfer into English
   6. Rearrangement
   7. Particle insertion
C. The printout

The GAT dictionary consists of 3,700 split and 2,370 unsplit entries. They are on two magnetic tapes, and their lengths are 155 and 175 characters, respectively. These record lengths contain the Russian entry, the paradigmatic and other grammatical information

used for subsequent machine analysis, translation codes, and one or more English equivalents.

The GAT-dictionary record lengths contain invariant information (inherently present in the Russian entry) and variant information (assignable to the Russian entry on the basis of the source text fed in). The latter information is generated by the computer.

The updating of the dictionary is a relatively easy procedure, since the increase in the dictionary entries is not directly tied up with the changes in the logical assembly.

The curve reflecting the ratio of the increase in dictionary entries as a function of the length of the corpus is shown on the following page in Figure 1.

The GAT dictionary is operationally used in two portions:

1. The split portion contains items which can be inflected.
2. The unsplit dictionary carries unsplit forms subdivided into several groups for specific reasons. These groups are listed in Appendix I to this paper.

If the Russian entry can have several English equivalents, a special code calls in the battery of tests to decide which equivalent is to be selected. The English equivalent, if inflected, is entered as a stem. The computer generates the necessary English form according to a code in the Russian entry.

The dictionary has been extracted in part from Russian scientific chemical texts and in part from metallurgical texts. The future development of the dictionary will proceed along the lines of building up a general (macro) dictionary, supplemented by field (micro) dictionaries. We envision, for example, that a text dealing with heart surgery will utilize three dictionaries: general, medical, and cardiological.

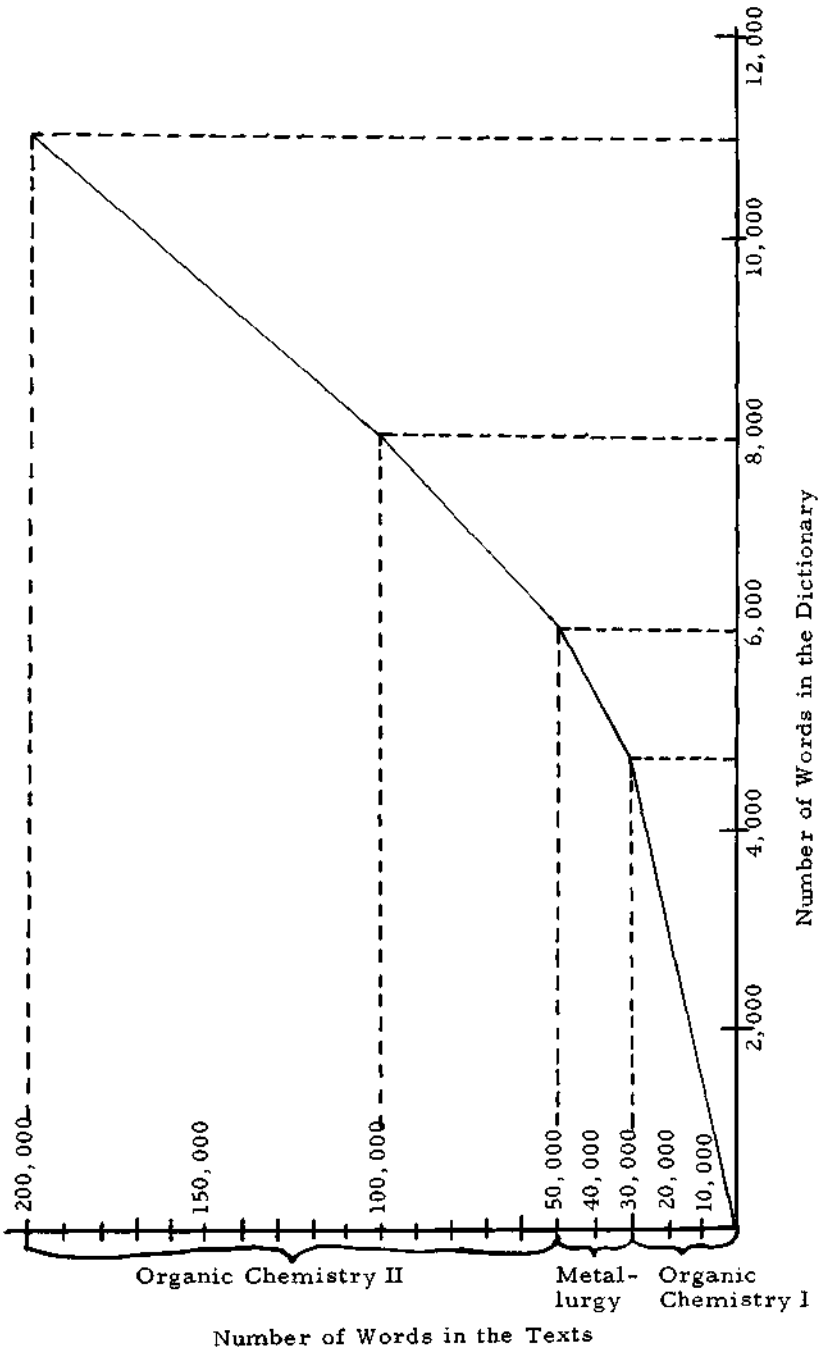
With the experience gained thus far, the further expansion of the dictionary should be carried out by means of extracting the equivalents on the basis of bilingual texts. We feel that in this manner a machine translation dictionary can be made operational that will contain relatively fewer multiple semantic choices.

The internal structure of the dictionary will be changed eventually for two reasons:

1. A more compressed format is desirable.

Figure 1

GAT TEXT-DERIVED DICTIONARY
Increment Curve



2. There is a need for introducing inherent semantic codes with each entry. These codes will serve later as a cue for the solving of some of the temporary ambiguities within the governing structures.

This dictionary is run on the IBM 705 and is currently being adapted for the RCA 501 and the Philco S-2000.

The corpora, analyzed and unanalyzed, now total approximately 180, 000 words in the field of organic chemistry, and 20, 000 words in the field of metallurgy.

The text is being further expanded. At the present time an additional 100, 000 words have been keypunched in the same field. The preceding chart seems to indicate that the lexical abstraction of some 500, 000 words of continuous texts may be sufficient to produce acceptable translation in the field of organic chemistry. The addition of two routines to the interpolation procedure will strengthen this probability. These are:

1. A procedure for generating some of the English chemical terms not in the dictionary
2. A procedure for identifying the word-class and grammatical features of items not in the dictionary, and their cases and numbers, if any

At the present stage, the problems under 1 and 2 are in a preliminary phase. They are merely listed by the computer as problems to be solved.

The runs on June 8, 1959, (when 100,000 words were translated on the IBM 705) disclosed a series of inadequacies in our program. They can be grouped in the following manner:

1. Inadequacies resulting from dictionary gaps.
   A sentence could not be handled if some 20% of the words were missing from the dictionary. This fact prompted us to start working on the two routines mentioned above.
2. Failures in the subject-predicate routines.
3. Failures in the insertion of English prepositions (e.g. , in front of a noun with preceding modifiers) and syntactic insertion such as "is, are/was, were" (if the modifiers in front of the word were present).
4. An accurate evaluation of the rearrangement operation was not possible because its failure to go into operation

was due not to intrinsic defects, but rather to the absence of the appropriate code in the preceding routines.

On June 8, 1959, we also ran a random text of about 1, 000 words. The failures described above were primarily noted and analyzed on the basis of this text. These are listed in detail in Appendix II to this paper. Altogether, the GAT as it stands now has approximately 30, 000 instructions, which perform approximately 40 complex operations. Out of 442 performances of these 40 operations, 590 failures have been registered, i.e. , 13.4%.

As far as the linguistic analysis is concerned, the algorithmic operations work on three levels:

1. The first level -- morphemic analysis -- is concerned with the analysis of the individual word. Any remaining ambiguities are resolved by examining the word class, number, and case of the surrounding words. After the word is assigned to a word class and its subclasses (e.g., case, number, animation, person, voice, tense) the second level of analysis goes into effect.
2. The second level -- syntagmatic analysis -- assigns a sequence of word classes to one of the three word-combination types (agreement, government, apposition) on the basis of the logical trees which indicate the permissible points of entry for a word combination, as well as the type of the word combination itself. In Appendix III the word-combination types (syntagmatic groups) are described in more detail.
3. The third level -- syntactic analysis -- defines the sentence as a concordance between a subject and its predicate, and the rest of the syntagmatic groups are treated as if they were either parts of the noun phrase or parts of the verb phrase.

There are two operations carried out prior to syntactic analysis proper:

1. The exclusion operation

   For the purpose of this operation, we define "exclusion" as a stretch of two or more words (items) within a sentence which, owing to specific circumstances, can be transferred directly or translated word-for-word from

the input to the output language. An exclusion stretch is normally bound by punctuation marks. Thus members of an exclusion group, 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, and certain other subclauses, such as preposition structures containing some formula.

2. Sentence separation routine
The sentence as fed into the computer from the raw text may be simple (noun-verb), complex, or compound. To facilitate the syntactic analysis, an operation is carried out to break down the compound and complex into simple sentences of the 00, 01, 10, 11 type.
The simple sentence in Russian has one noun phrase and one verb phrase; this we label a 1-1 type. The first digit has been arbitrarily chosen to represent H (the head word of a noun phrase); the second, P (the head word of a verb phrase), although this is not connected with order. More than one H and P is called a 2-2 type, where either component may be two or more in number.

The insertion of the article follows. The next operation is that of rearrangement, followed by the printout. In Appendix IV, samples of our printouts, accompanied by Russian texts, are given.

The primary criterion by which we are guided in analyzing the output is the degree of accuracy with which the source information has been transferred into the target language.

On January 25, 1960, new samples were run.

- - -

In addition to the GAT from Russian to English, there are other objectives pursued by our project.

Trans-Slavic

The purpose of the basic research in the field of multi-Slavic structures is the design of a common program for morphological analysis of Slavic languages. Instead of analyzing each language separately, the algorithms are being organized in such a manner

that they will permit an individual morphological analysis of Russian, Czech, and Serbo-Croatian, by using identical rules to the extent possible. This research will be the subject of a future paper.

Chinese and Arabic

Preliminary efforts are directed toward the preparation of algorithms for translation from English into Chinese and Arabic.

The Chinese section has begun the work of setting up form classes for Chinese. Study has been concentrated on the various possible translations of some English sentences into Chinese sentences. The purpose is to establish a complete and original analysis of Chinese grammatical classification in order to engender certain basic rules for Chinese translation. Since there is no alphabet in the Chinese language, the Chinese Standard Telegraphic Code will be used for the 4-5,000 most commonly used Chinese characters.

The coding of the United Nations Charter has been completed.

The Simulated Linguistic Computer

The work in French-to-English translation at Georgetown has been colored by the simulation system that is used to mechanize the linguistic rules. The programming system requires a fairly large expenditure of effort and memory space on the interpreting routines; but whenever an individual linguistic operation is added to the system, this takes only a small extra amount of space.

Thus the over-all translation system has evolved as a large number of small and medium-sized routines for handling specific problems. Any routine can be called into action at any time during the translation process, and it can be called in more than once for a single sentence. Hence it is not possible for us to describe the French-to-English system as a fixed series of processes, each one achieving a well-defined part of the conversion from one language to the other. Of course, the dictionary lookup is a separate and distinct operation, and it is always the first step in the translation. We think most machine translators are agreed, if on nothing else, on consulting the dictionary only once per cycle, at least until some new kind of large-scale memory is available.

The lookup is done on batches of about 1,200 words of input text. This limited size was chosen in order to eliminate any need for tape sorting and merging on an IBM 704 computer with 8, 000 words of

core storage and 8, 000 words of drum storage. As the computer on which, the programs were developed has acquired a 32, 000-word memory, the program will eventually be modified to process batches of about 2, 000 words of text. Its rate of translation should increase from the present 10, 000 words an hour to at least 12, 000.

After the dictionary has been consulted, the French words that could not be found in it are matched against a list of common word endings. When a French word-ending is found to be in the table, the table gives an assumed grammatical code for the word, and an English word-ending to replace the French one in the transliteration of the unknown word. For instance, the word polymeriserait could not be found in the dictionary as it stands, but the ending -iserait would be identified in this list, and the word would come out in the final translation as "would polymerize". This kind of modified transliteration would not give such plausible results when applied to Russian, but the deduction of the probable grammatical value of an unknown word from its ending is even more important in Russian, to prevent the sentence structure from collapsing.

This guessing at unknown words is the last operation that occurs in a fixed sequence, until the moment for outputting the translation arrives. The time in between is taken up in obeying instructions according to the priority numbers they contain. Whenever an instruction is obeyed, the linguistic operation which it names is located, and the first command in the operation is carried out. This is the point at which the name "simulated linguistic computer" becomes appropriate. The linguistic operation consists of a group of constants plus a series of commands. A command consists of an operation code, a data address, a count, two control-transfer addresses, and what amounts to an address for the item in the sentence that is to be tested or altered.

Throughout the translation routine, the items in the sentence are held in memory in such a way that the macro-orders can look on them as occupying locations in a specially structured memory. There are 128 locations available, and those that are occupied by items in a sentence form a list structure in which it is possible to begin at location zero, move either rightward or leftward through the whole sentence, and arrive at location zero again. This is a

ring-shaped list of limited size, with item zero lying just before the the first item, and just after the last one. Every location in this system, which could be called the simulated linguistic computer's memory, includes fixed-length stores for grammatical coding, flag bits, and links to its right-hand and left-hand neighbors, plus variable-length storage for English equivalents, instructions, and secondary grammatical codes.

The item in location zero, besides showing where the ends of the sentence are, supplies a permanent set of instructions to every sentence, regardless of the words it contains. These general instructions do provide a certain framework for deciding on the timing and sequencing of ordinary linguistic instructions. In order of performance, they have the following functions:

First, the resolution of grammatical ambiguity in le, la, les, en and the rather large number of third-person singular, present tense, verb, forms that are homonymous with nouns. In certain cases, the right answer may not be arrived at here, but at least the system allows us to insert any specific solution to these problems at any later point in the sequence.

Second, the addition of inflectional suffixes to words in the expected English output. The regular verbs and nouns are handled by the general operation, and so are the verbs "to be", "to have", "to do", "can", and "must". Other irregular verbs, as well as irregular nouns, get their inflections through individual instructions just before or just after this general instruction is executed. Simple negation and the interrogative inversions in English are handled as part of the process of inflection.

Third, adjectives are shifted from their usual French position in the noun phrase to the English position.

Fourth, the English indefinite article is changed from "a" to "an" before words beginning with vowels.

This seems like a very sketchy list of operations for translation. There are actually several other general operations in the system, but their functions are very minor; they are called "general" simply because it is convenient to supply them with every sentence rather than with particular words in the dictionary. The great majority of the work is done by instructions found in the dictionary, so that

it is not possible to say what instructions will be executed in translating a sentence until the sentence is known.

Besides the general operations, there are quite a large number of operations of wide applicability. For instance, many French verbs when used in the reflexive must be translated by the English passive, and many more require the English active. There are two standard instructions for inclusion in the dictionary entries of verbs of these two types. Then there are many verbs which have one English equivalent when used in the active, and a different English equivalent, usually an active verb, when they are used in the reflexive in French. There is a standard instruction for this, but every time it is put into a dictionary entry, it has to be accompanied by the particular English equivalent that it will substitute when the verb turns out to be part of a reflexive construction. The English equivalent could be called a parameter for the instruction. Or in programming terms, a linguistic operation is coded as a fairly short program for the simulated linguistic computer; this program is initiated by a calling sequence consisting of at least an instruction, plus as many parameters as may be needed.

The French-to-English dictionary and operations reached approximately their present condition last June. The text on punched cards consists of about 230, 000 words. The first 200, 000 words of this text made up the examined corpus, of which the first 20, 000 words have now been run on the IBM 704, and the last 30, 000 were left unopened in their boxes until this year. This was done partly because 200, 000 words seemed like more than enough to worry about at once, and partly in order to have a random unprepared text available which would still be similar in character to the examined text. The first 5, 000 words or so of this random text have now been run, with results that can be inspected. Apart from the mistakes due to faulty keypunching, we do not find it possible to name a small number of categories into which most of the translation errors fall. There must be at least 100 linguistic operations that will need revision when we find the causes of the mistakes, and no doubt 100 new operations will have to be coded.

With a view to making the simulated linguistic computer system a more convenient vehicle for our Russian-to-English work we are

adding a number of improvements to the program. One of them will double the size of the fixed-length grammatical area in each item location. Thirty-six bits offers plenty of elbow room for French, but the complexity of Russian noun and adjective morphology makes it very desirable to allow a separate bit position for every combination of case and gender, counting plural as a fourth gender. Another addition to the system should make it simple to mark and identify a large number of different chains within the sentence.

An SLC program has been written for the IBM 709 computer and is now being checked out. It embodies the improvements being made to the IBM 704 program, and in any case it will handle the existing French dictionary and operation material exactly as the IBM 704 program does now. The system has also been programmed for the Philco S-2000, but almost nothing has been done toward checking it out.

The differences between general computer programming and the sort of linguistic programming that the simulation system will handle have made us doubtful about the necessity or value of a symbolic programming system for the SLC. However, we have recently written a description of a symbolic language which might be helpful in linguistic programming and which we think could be handled by a much simpler assembly program than, for example, the SAP assemblers. It would be less ambitious than COMIT, since the symbolic operations would pass through three stages. First, the symbolic decks would be read by the assembly program and converted into files of absolutely coded material, in the same format as the system accepts now. Second, the file of operations would be read by a routine in the translation program and incorporated into that program, which would then write itself out as a self-loading program on tape. And third, the lookup and translation programs would translate. The second and third stages are what already exist in the system; the symbolic-to-absolute conversion would be a separate stage that would not burden the translation system at run time.

APPENDIX I

UNSPLIT DICTIONARY (p. 1-36)

1 Abbreviations (B., CAS.)
2 Formulae (C$U)
3 Numerals cardinal (-10, 122)
4 Formulae (G1, G16$P) Greek letters
5 Single Latin letters
6 Punctuation marks
7 Signs
3 Short adjectives masculine of two stems
4 Adverbs
5 Prepositions
6 Conjunctions
7 Particles
11 Nouns, masculine, (INOSTRANEQ); form, which stem is not retained in most cases
13 Nouns, neuter, (CISEL); form, which stem is not retained in most cases
12 Nouns, feminine, (BANOK); form, which stem is not retained in most cases
14 Pronouns (NEH, NEMU, ON, ONA)
16 Pronouns with noun forms (NAM, NAMI)
21 Form of "to be"
24 Gerund
26 Past of "to be" plus irregular high frequency items
30 Numerals (DVA, TRI)
32 Possessive pronouns (SAMYX, SVOE)
35 Definite demonstrative pronouns
36 Personal pronouns without gender, reflexive
37 Numerals ordinal and collective
45 Pronouns (El)
145 Pronouns (EMU, IMI)
142 Short adjectives, neuter only
1202 Degree
Idioms 1-49-57

APPENDIX II

STATISTICS ON THE RANDOM CORPUS

	Operations		Total	Correct	Failure	%
1	Word lookup		1530	1418	112	7.3
2	Idiomatic		14	11	3	21.0
3	Interpolation		388	300	88	22. 6
4	Apposition		12	12	-	0.0
5	Agreement		190	133	57	30.0
6	Noun Government		129	93	36	28.0
7	Preposition Government		113	87	26	23.0
8	Adverb Government		-			
9	Adjective Government		-			
10	Participle Government		21	11	10	49.5
11	Sentence Recognition	219				
a	Sentence Type	00	58	52	6	10.3
b	Sentence Type	01	34	22	12	35.3
c	Sentence Type	10	58	52	6	13.6
d	Sentence Type	11	37	29	8	21.6
e	Sentence Type	12	1	1	0	0.0
f	Sentence Type	20	8	7	1	12.5
g	Sentence Type	21	13	7	6	46.0
12	Lexical Choice		114	96	18	15.1
13	Synthesis		337	336	1	0.3
14	Rearrangement		52	27	25	43.0
15	"OF" Insertion		139	86	53	30.8
16	"BY" Insertion		18	15	3	17.0
17	"TO" Insertion		2	2	-	0.0
18	Subject		125	85	40	32.0
19	Singular-Plural		336	336	-	0.0
20	Past Tense		45	45	-	0.0
21	Present Tense		29	29	-	0.0
22	Future Tense		-			
23	Participle		57	51	6	10.4
24	Predicate		82	73	9	11.0
25	English Article Insertion		not studied			

APPENDIX II (Continued)

	Operations	Total	Correct	Failure	%
26	Exclusion Operation	not studied			
27	Adjective-Nouns	6	6	-	0.0
28	Case Ambiguity of Nouns	90	66	24	26. 6
29	Case Ambiguity of Adjectives	108	98	10	9.2
30	Syntactic Insertion (if, they, there)	16	11	5	31.0
31	Morphological Analysis	284	283	1	0. 3
32	Number Ambiguity Resolution	49	43	6	12.2
41	Verb Government	26	8	18	69.5
		4421	3831	590	13.4

APPENDIX III

Midway between the morphological analysis of single words taken separately, and the syntactic analysis of whole sentences to determine the head words in their constituent noun and verb phrases, stands syntagmatic analysis, conducted within either a noun phrase or a verb phrase on certain groups of adjacent words related by form or meaning. For the purpose of assigning computer codes, a syntagmatic series may be defined as a combination of at least two words, of which one may be designated as a "pivot word", in order to create a unit for which the pivot word alone may be substituted. Either of the two components of a syntagmatic series may consist of more than one word, linked by punctuation or a mediating word, usually a conjunction or particle; but the series, no matter how great the expansion of its component parts, is always reducible to a pivot element and an element in apposition to, governed by, or in agreement with it. When adjacent adjectival or adjacent nominal words have the same morphological codes, they are assigned a special presyntagmatic code of 3001, 3002, 3003, 3004, 3005, or 3006, if they are adjectives without ambiguities (one case, one gender, one number only). If they are ambiguous on any level, then the code is HOMO. The code HOMO is used when two or more nouns mirror each other completely. The special code indicates that the words share a homogeneous function in the nominative, genitive, dative, accusative, instrumental, or prepositional case. This preliminary linking of adjectives with the same form and function shortens and simplifies subsequent syntagmatic operations. Following homogeneous-function analysis, the computer searches for three different types of syntagmatic unit: apposition, agreement, and government, in that order. The following paragraphs will give brief descriptions of the computer codes attached to these syntagmatic types, along with the distinguishing features of each.

A. Apposition

Apposition structures, as defined for mechanical translation, are combinations of an uninflected word, an adverb, with another inflected or uninflected word, in order to form a meaningful

unit determined not by morphologically expressed case relationships, but simply by contiguity and semantic factors. Juxtapositions of more than one nominal element sharing the same referent, generally classified as apposition structures in English grammar, are not subject to computer classification as a special syntagmatic type, since their correct translation may be effected by homogeneous-function codes attached to nominal words within government and agreement structures.

Although Russian grammar defines as apposition structures combinations of verb plus complementary infinitive, verb plus gerund, or verb plus uninflected comparative adjective, computer syntagmatic apposition analysis is confined solely to adverbs and other word classes with which they can enter into meaningful relationships.

The computer check for an appositional relationship is a search for that part of speech immediately preceding and that immediately following an adverb. If no verb, adjective, or adverb is adjacent to the adverb being analyzed, the latter is considered outside the scope of syntagmatic analysis and is assigned codes in a subsequent syntactic operation, outlined in the paper on Syntactic Analysis.

In the codes assigned to each of the elements in apposition structures, the first digit is always 4, denoting an adverb; the second digit may be 2, 3, or 4, according to whether the adverb is in a meaningful relationship with a verb (2), adjective or participle (3), adverb or gerund (4); the third digit is always a 3, designating an apposition structure; and the following P or F indicates that the second element in the series precedes or follows the adverb being analyzed.

1. The most frequently encountered apposition structure is the combination of an adverb with a verb or gerund:

NEREDKO	SLUCALOS6	"it frequently happened"
423F	423F	
VOOB5E	GOVOR4	"generally speaking"
443F	443F	

Here the verb "happened" and the gerund "speaking" are considered the pivot words, to which the adverbs "frequently" and "generally" are in apposition.

2. An adverb with an adjective or a participle is considered an apposition structure:

NAIBOLEE	DREVNI1	"most ancient"
433F	433F	
DALEKO	ZAWEDWI1	"far advanced"
433F	433F	

Here the adverbs "most" and "far" are in apposition to the adjectives "ancient" and the participle "advanced", which are considered the pivot words in the syntagmatic series.

3. An adverb in combination with another adverb forms an apposition structure, the pivot word depending on the information communicated:

POCTI	POSTO4NNO	"almost constantly"
443F	443F	

In this utterance "constantly" is considered the pivot word, as it may replace the whole syntagmatic series.

B. Agreement

An agreement structure exists where an adjectival word linked to a nominal word agrees with it, that is to say, is assigned the same gender, number, and case indicators as the nominal word. Agreement analysis considers as nominal words nouns, personal and reflexive pronouns, adjectives with noun function, and cardinal numbers in the nominative and accusative cases; adjectival words include adjectives, participles, non-personal and non-reflexive pronouns, and ordinal numbers. In the case of Russian third-person possessive pronoun adjectives, where there is no formal case differentation, such words are regarded as sharing the case of their pivot word.

Of the 12 types of agreement possible in Russian, only 6 are treated for computer purposes in syntagmatic analysis. Any instance of agreement between a word in the noun phrase and a word in the verb phrase remains unanalyzed until the computer syntactic operation goes into effect.

To all members of an agreement structure a four-digit recognition code is assigned. The first digit is the number 3, signifying the presence of an adjective; the second is the number 1, denoting a noun; the third, the number 1, representing an agreement structure; and the fourth, any number from 1 through 6, 1 for nominative, 2 for genitive, 3 for dative, 4 for accusative, 5 for instrumental, and 6 for prepositional case:

IX	KATALITICESKOMU	DE1STVIH	"their catalytic
3113	3113	3113	action"

In this construction the adjectives "their" and "catalytic" are both assigned codes designating an agreement structure with the noun "action" in the dative case, even though the Russian equivalent of "their" has no formal case marker.

Of the 1,868 agreement structures coded for the examined text, 714 were in the genitive case, 563 in the nominative, 205 in the prepositional, 170 in the accusative, 158 in the instrumental, and 58 in the dative. The range of expansion of the pivot word to the left and right in the syntagmatic series of the examined text was 5 words to the left and 5 to the right for syntagmatic units in the nominative case, 7 left and 4 right for those in the genitive, 6 left and 2 right for dative, 5 left and 3 right for accusative, 6 left and 3 right for instrumental, and 4 left and 3 right for the prepositional case. Isolated instances where the pivot word was expanded as much as 18 words to the left and 17 to the right were not assigned computer codes. In the prevailing majority of instances the pivot word in the agreement structure combined with one or two adjectives, with 28 cases of three agreeing adjectives and 2 cases of four agreeing adjectives also occurring in the examined text.

In addition to defining agreement structures, the agreement code is also used for unloading government codes, detecting the subject of the sentence, inserting English articles into the Russian text, introducing English prepositions to translate certain Russian case relationships, establishing the heads of the noun phrase and verb phrase participating in nucleus structures, and effecting word-order rearrangement.

C. Government

In a government structure the relationship existing between a nominal word (that is, a noun, personal or reflexive pronoun, cardinal number in nominative or accusative case, or adjective used as a noun) and another inflected or uninflected word, usually preceding it, specifically determines a particular non-nominative case in the nominal word. The governing word in such structures is regarded as the pivot word.

The computer analyzes a government structure by searching for the part of speech preceding and following the item analyzed. For all

members of a government structure a four-digit numerical code is assigned. The first digit may be a number from 1 to 5, showing that the first member of the structure is a noun, verb, adjective, adverb, or preposition, in that order; a gerund is coded as 4, a participle as 3. The second code digit may be a 1, indicating that the governed element is a noun, 3 if it is an adjective with noun function, 5 if a preposition occurs as the intervening element in a weak verbal government structure. The third digit, always a 2, classifies the syntagmatic unit as a government structure. Finally, in the last position, there may be a number from 2 to 6, indicating whether the governed element is in the genitive, dative, accusative, instrumental, or prepositional case. A terminal 0 may also occur in a government-structure code, identifying it as a weak government structure.

1. Nouns may combine with verbs in strong or weak government structure:

In a strong verb-government structure the governed noun displays a case as determined by the verb case-determiner.

PROFESSOR	DAL	OPREDELENIE	"The professor
	2124	2124	gave a definition"

The relationship expressed by the transitive verb followed by a direct object in English is rendered in Russian by the verb plus the accusative case; that is to say, the Russian verb in this utterance governs the accusative case in the noun object. The last digit in a strong verbal government code can never be a 6, as the prepositional case may follow only a preposition, which does not participate in strong verbal government.

In a weak verbal government structure a verb cannot be linked with a following noun object without the intermediary of an intervening preposition, dependent itself on the preceding verb, but at the same time determining the case of the noun following it. Weak verbal government is considered distinct from prepositional government, where the preposition does not function as a mediating word between a verb and its object. To mark a weak verbal government structure, the computer assigns one code shared by the verb and following preposition, another shared by the preposition and its object. The verb-preposition code is 2520, indicating a verb followed by ft preposition in a weak government structure. The preposition-object code is 512x (the last digit varying from 2 to 6 according to

the case of the object), designating a government preposition-object relation:

OBRA5AHT NA SEB4 VNIMANIE "attract attention"
2520 2520
5124 5124 5124

In the Russian expression the verb governs the noun weakly, through the intermediary of a preposition (which in turn governs the noun strongly, i. e. , directly). The fact that the preposition in this particular translation has a zero English rendering is meant to emphasize the point that the structural coding is based on the Russian relationships, and that the English translation is later determined by that coding.

2. Prepositions always determine a specific case in following nominal objects. Where a given preposition may have a different translation depending on the case of its object, a check of the morphological codes assigned the object will resolve ambiguities in translation. This operation is described in greater detail in the paper on Lexical Choice.

POSLE OPERAQII "after the operation"
5128 5122

The object of the Russian preposition here is in the genitive case--this particular preposition in this utterance governs the genitive case in its object.

3. Nouns may govern a specific case in an immediately following noun:

ORGAN ZRENI4 "organ of sight"
1122 1122

Here the function assumed by the English preposition "of" is taken by the genitive case of the governed noun in Russian.

4. Certain adjectives are case determiners:

POLNI1 STRAXA "full of fear"
3122 3122

In this Russian construction the genitive case once more assumes the function of the English preposition "of".

5. Gerunds may enter into government structures:

DELA4 ZAMETKI "making notes"
4124 4124

The Russian gerund, rendered in English by "making", governs the accusative in the following noun "notes".

6. Certain adverbs, especially those of comparative degree, participate in government structures:

ON	3TO	LUCWE 4122	MEN4 4122	ZHAET	"He knows that better than I".

The English phrase "better than I" is rendered in Russian by a comparative adverb governing the genitive case of the following personal pronoun.

The three syntagmatic types just described may all be combined to expand a single pivot word, as the following example illustrates:

V 5126	DALEKO 5126 433F 3116	ZAWEDWIX 5126 433F 3116	SLUCA4X 5126 3116 1122	GLAUKOMY 5126	"in far-advanced cases of glaucoma".

In the Russian utterance the pivot word "cases" is in an agreement structure with "advanced", which in its turn is involved in an apposition relationship with "far"; "cases" also governs the following noun "(of) glaucoma". The maximum expansion by all syntagmatic types of a pivot word encountered in the examined text was eleven words.

Punctuation marks such as commas and dashes are important in syntagmatic analysis not as markers for intonational patterns of pause or change in pitch, but as guides in locating so-called loops. A loop may be defined as any set of words modifying an element in an Utterance without changing the noun-phrase--verb-phrase structure of that utterance. According to this definition a syntagmatic series may be a part of the whole of a loop. Commas and dashes frequently demarcate loops, but they also serve as separators of noun-phrase--verb-phrase utterances. Syntagmatic analysis checks the words on either side of a comma or dash for indicators of membership in a single syntagmatic series, in order to establish whether groups of words bounded by such punctuation are loops modifying a word in one utterance, or parts of two different utterances.

Since a syntagmatic series is replaceable by its pivot word, it may be seen that syntagmatic analysis performs the valuable service of reducing extensive strings of words to units which may be transposed during word-order analysis in order to achieve a more English sentence than would result from a word-for-word translation from the source language. For example, in a sentence translated without rearrangement from the Russian as: "Together with this in

the reaction of oxidations became noticeable certain additional influences", syntactic analysis will produce codes for "together with this", "in the reaction of oxidation", and "certain additional influences", permitting them to be rearranged to produce a smoother English sentence: "Together with this certain additional influences became noticeable in the reaction of oxidation".

Syntagmatic units may also be used in another level of text analysis. From a sentence of the text input -- bounded by sentence separators such as periods, commas, spaces, question marks, and exclamation marks--the computer extracts all the component nucleus structures, those combinations of a personal marker (verb) and a word in the nominative case, or substitutes for such combinations.

APPENDIX IV

Russian Text

3KSPERIMENTAL6NA4 CAST6 ▭

V REZUL6TATE KATALITICESKOI KONDENSAQII S METILOVYM

SPIRTOM NAD AKTIVIROVANNOI SOL4NOI KISLOTOI GLINOIOGUMBRINOM

POLUCENY DVUSLOINYE KONDENSATY *

VODNYI I MASL4NYI SLOI RAZDEL4LIS6, OT KAJDOGO SLO4 OTGON4LS4

PRODUKT , KIP45II DO 100 @ *

MASL4NISTYI OSTATOK , KIP45II VYWE 100 @ , MNOGOKRATNO

3KSTRAGIROVALS4 10 % N$AOH *

POLUCENNYE FENOL4TY IZVLEKALIS6 3FIROM DL4 OSVOBOJDENI4 OT CASTICNO

UVLECENNYX 5ELOC6H NEITRAL6NYX MASEL I RAZLAGALIS6

RAZBAVLENNOI H#2SO#4 * ▭

POSLE VYSUWIVANI4 3FIRNOGO RASTVORA I UDALENI4 3FIRA

POLUCALAS6 SMES6 FENOLOV , KOTORA4 ZATEM PODVERGALAS6

FRAKQIONIROVANIH *

DO 182 @ OTGON4LS4 NEPROREAGIROVAVWII FENOL *

DALEE NA KOLONKE FRAKQIONIROVALIS6 PROIZVODNYE FENOLA *

APPENDIX IV (Continued)

FENOLY ANALIZIROVALIS6 METODOM ARILIROVANI4 B# /1/5 E# *

POLUCENNYE KRISTALLICESKIE ARILGLIKOLEVYE KISLOTY RAZDEL4LIS6 DROBNOI

KRISTALLIZAQIE1 * □

V PROQESSE RABOTY BYLO USTANOVLENO , CTO KATALIZATOR POSLE 18OCASOVO1

RABOTY TER4L SVOH AKTIVNOST6 POCTI NA 30 % *

□□ 003

TAKA4 DEZAKTIVAQI4 KATALIZATORA VYZYVAETS4 , VERO4TNO ,

OTLOJENI4MI SMOLISTYX OSTATKOV I UGL4 NA EGO POVERXNOSTI , V

REZUL6TATE CEGO POVERXNOST6 KATALIZATORA STANOVITS4

NEDOSTUPNO1 DL4 REAGIRUH5IX VE5ESTV *

VO VSEX OPYTAX POSLE 6OCASOVO1 RABOTY KATALIZATOR

REGENERIROVALS4 PUTEM PRODUVANI4 CEREZ

REAKQIONNUH TRUBKU VOZDUXA PRI TEMPERATURE 350 @ *

POSLE 18OCASOVO1 RABOTY V REAKQIONNUH TRUBKU POME5ALAS6 SVEJA4

PORQI4 KATALIZATORA * □

ISXODNYMI VE5ESTVAMI SJUJILI SVEJEPEREGNANNY1 VYSUWENNY1

FENOL I PEREGNANNY1 METILOVY1 SPIRIT * □

APPENDIX IV (Continued)

English Translation

AN EXPERIMENTAL PART

AS A RESULT OF CATALYTIC CONDENSATION WITH METHYL ALCOHOL OVER ACTIVATED HYDROCHLORIC ACID BY A GUMBRIN CLAY THEY ARE OBTAINED TWO-LAYERED CONDENSATES. AQUEOUS AND OILY LAYERS RAZDEL4LI, FROM EACH LAYER THERE WAS DISTILLED A PRODUCT, WHICH BOILS UP TO 100-. THE OILY RESIDUE, WHICH BOILS ABOVE 100-, REPEATED WERE EXTRACTED 10(N$AOH. OBTAINED PHENOLATES WERE EXTRACTED WITH ETHER FOR LIBERATION FROM PARTIALLY INVOLVED BY THE ALKALI OF NEUTRAL OILS AND DECOMPOSED DILUTED H=2SO=4.

AFTER THE DRYING OF AN ETHER SOLUTION AND THE REMOVAL OF THE ESTER THERE WAS OBTAINED A MIXTURE OF PHENOLS, WHICH THEN WAS SUBJECTED TO FRACTIONATION. UP TO 182- THERE WAS DISTILLED NON-REACTED PHENOL. FURTHER THROUGH A COLUMN THERE WERE FRACTIONATED DERIVATIVE OF PHENOL. PHENOLS WERE ANALYZED BY A METHOD ARILIROVANI4 //15//. OBTAINED CRYSTALLINE ARYLGLYCOLIC ACIDS RAZDEL4LI FRACTIONAL BY CRYSTALLIZATION.

IN THE PROCESS OF WORK WAS ESTABLISHED, THAT A CATALYST AFTER 18-CASOVO1 WORKS LOST ITS ACTIVITY ALMOST ON 30(. SUCH DEACTIVATION OF A CATALYST IS CAUSED, PROBABLY, BY THE DEPOSITS OF THE TARRY RESIDUES AND COAL ON ITS OF SURFACE, AS A RESULT WHICH THE SURFACE OF A CATALYST BECOMES INACCESSIBLE FOR REACTING SUBSTANCES. IN ALL EXPERIMENTS AFTER 6-CASOVO1 WORK A CATALYST WAS REGENERATED BY MEANS OF PURGING THROUGH THE REACTION TUBE OF AIR AT A TEMPERATURE 350-. AFTER 18-CASOVO1 OF WORK INTO A REACTION TUBE WAS PLACED THE FRESH PORTION OF A CATALYST.

BY THE INITIAL SUBSTANCES THERE SERVED FRESHLY DISTILLED DRIED PHENOL AND DISTILLED METHYL ALCOHOL.