UNCLASSIFIED

AD 402 242

Reproduced by the

DEFENSE DOCUMENTATION CENTER

FOR

SCIENTIFIC AND TECHNICAL INFORMATION

CAMERON STATION, ALEXANDRIA, VIRGINIA

UNCLASSIFIED

RADC-TDR-63-102

JANUARY 15, 1963

63-3-2

FINAL REPORT FOR **Automatic Language Translation**

INTERNATIONAL BUSINESS MACHINES CORPORATION
IBM RESEARCH
Thomas J. Watson Research Center
Yorktown Heights, New York

CONTRACT AF30(602)-2617

Prepared for
INFORMATION PROCESSING LABORATORY
RESEARCH AND TECHNOLOGY DIVISION
ROME AIR DEVELOPMENT CENTER (AFSC)
Griffiss Air Force Base
New York

$15.50

(13) RADC-TDR-63-102 — 765700 (5)-(872300) (9) JAN[redacted] (15) [redacted]63

(7) FINAL REPORT, ~~FOR~~ (16) Automatic Language Translation

(8) N.A.
(10) 224p. incl. illus-tables
(11) N.A.

INTERNATIONAL BUSINESS MACHINES CORPORATION
IBM RESEARCH
Thomas J. Watson Research Center
Yorktown Heights, New York

(12) **CONTRACT AF30(602)-2617**

J.C.

Prepared for
INFORMATION PROCESSING LABORATORY
RESEARCH AND TECHNOLOGY DIVISION
ROME AIR DEVELOPMENT CENTER (AFSC)
Griffiss Air Force Base
New York

FOREWORD

This report summarizes and documents the activities in improving the capabilities of the AN/GSQ-16 language translation complex performed under contract AF 30(602)-2617. The objectives of this contract were to expand further the linguistic capabilities of the AN/GSQ-16 (Mark II) system by developing a fully operational single-pass bi-directional translation system, and by augmenting the multipass dictionary to satisfy the long-range goals as well as to perform system studies to increase the ability and effectiveness of the complex.

Both the bidirectional translation system and the expanded multipass dictionary were demonstrated during the course of the year. The results of the system studies are included in this report.

The lexicographic effort of this project were greatly aided by the assistance of the Aerospace Information Division of the Library of Congress and the MT group of Syracuse University.

The work at IBM Research was carried out under the direction of Dr. Gilbert W. King and Dr. Ernest H. Goldman.

ABSTRACT

This report describes the accomplishments for [redacted] 1962 in the development of the AN/GSQ-16 language translating complex are described. ~~to provide~~ a greater degree of sophistication in machine translation. This involved studies in linguistic research, conversion to a bidirectional single-pass dictionary mode, and, in its incipient stage, to a multipass mode.

The search and programming techniques developed during the year, particularly the rapid subtable searching of core memory and the program for generative grammars, are described along with the basic linguistic researches into machine translation problems, compilation of data for the grammatical feature of government in Russian grammar, and generation of new grammar tables embodying grammatical analysis and sentence structure determination.

Improvements made, and those contemplated, to machine components and to the system organization are also reported.

TABLE OF CONTENTS

Section	Title	Page
1	Introduction	1
1.1	General	1
1.2	Background in Machine Translation	1
1.3	Description of Tasks	3
2	Discussion	7
2.1	Applied Linguistics	7
2.2	Linguistic Research	67
2.3	System Organization	173
3.	Conclusions and Recommendations	189
3.1	General	189
3.2	Applied Linguistics	190
3.3	Linguistic Research	191
3.4	System Organization	193
3.5	Photostore Improvements	194
	Appendices	197
Appendix I	Samples of Bidirectional Single-Pass Translation	198
Appendix II	Samples of Multipass Translation	217

LIST OF ILLUSTRATIONS

Figures	Title	Page
2.2-1	Diagram of a Russian Sentence in terms of a Simplified Grammar	73
2.2-2	Diagram of a Conventional Russian Grammar	74
2.2-3	Format for the Recording of Node Information	78
2.2-4	Diagram of Types of Nominal Government	110
2.2-5	Nominal Constructions with the Genitive	113
2.2-6	Relativity, Aspect, and Usage of Verbs	127
2.2-7	Subject Preference in Predication	129
2.2-8	Impersonal Constructions of the Verb	130
2.2-9	Direct Nominal Government Capabilities	131
2.2-10	Suggestive Usage of Nouns in Instrumental and Dative	134
2.2-11	Tentative Semantic Classes of Russian Nouns	136
2.2-12	Grammar Analysis Questionnaire	139
2.2-13	Graphic Representation of Constituent-Structure Rules	147

LIST OF ILLUSTRATIONS - Continued

2.2-14	Sentence Derivation Mapped into the Tree Structure for Rules of Formation	148
2.2-15	Derivation of the Sample Kernel Sentence	168

TABLES

No.	Title	Page
2-1	Search Input State Statistics for Multipass Translation of Russian Sentence to English.	179
2-2	Intermediate Passes - Sentence Analysis Statistics	180
2-3	Final Pass Statistics	181
2-4	Entry Statistics for Search Operations	182

Section I: INTRODUCTION

1.1 General

This report is intended to document the improvements in Automatic Language Translation accomplished during the past year under contract AF 30(602)-2617, describing in some detail the current status of the Russian-English MT dictionary, the associated language processing procedures and methodology, and the organizations developed for a next generation of the AN/GSQ-16 complex.

The objectives of this contract have been to advance the development of the AN/GSQ-16 language processing system primarily through improvements in the quality of translation. The effort has involved developing a final organization for the Russian-English MT dictionary and refining the syntactic and semantic analysis technique, while exploiting the earlier developments in the Rome Air Development Center program in machine translation. Specifically, the work has been aimed at bringing the MT dictionaries to operational status and at developing an organization for a production system.

1.2 Background in Machine Translation

The development of an automatic machine translation system, undertaken by the IBM Thomas J. Watson Research Center under contract to the U. S. Air Force Intelligence Processing Laboratory, Rome Air Development Center, has involved a systematic program of research and development in linguistic theory, language processing techniques, and system design. A description of the evolution of automatic language translation from the AN/GSQ-16 Mark I language processing system to the Mark II, and of the work performed to bring the Mark II system to its present successful operational status is given in the Final Report on Computer Set AN/GSQ-16 (XW-2), Contract AF 30(602)-2080, prepared for the Rome Air Development Center, Griffiss Air Force Base, Rome, New York.

Concurrent with the development of the Mark II language processing complex, IBM Research has conducted an extensive program in lexicon preparation to provide the system with a vast compilation of Russian stems, endings, full forms, and idiomatic phrases, and their English equivalents. This lexicon has been extensively evaluated and improved through the processing of several million words of Russian text.

In keeping with the evolutionary approach of the Air Force to machine translation, IBM Research has developed two translation capabilities for the Mark II system: a new and operational bidirectional single-pass translation mode, representing a transitional language process between the former unidirectional single-pass mode described in the final report of the Mark I system, AN/GSQ-16 (XW-1), and the new multipass Sentence Analyzer mode developed under contract AF 30(602)-2072 and described in the final report RADC - TDR - 62-105.

Evaluation on a fairly substantial corpus has demonstrated a marked improvement of this mode of translation over the unidirectional mode. Accordingly, during the performance of this contract, the entire IBM high-capacity MT operational dictionary has been converted to the bidirectional format and has been randomly tested on large volumes of text. The results indicate that a useful translation function has been achieved which will more than adequately serve, in an interim capacity, until the full capabilities of the multipass Sentence Analyzer mode are realized.

The program of research and development in language processing techniques has resulted in significant improvements in the accuracy, smoothness, and readability of automatic translation, and has provided a sound base for continued development.

This report describes the tasks that were carried out in the performance of contract AF 30(602)-2617 in connection with the development and construction of the Russian Master dictionaries for the bidirectional single-pass program, the investigation of grammatical analysis routines to be used in automatic translation of Russian to English, the linguistic research into machine translation, and the studies of improvements to the system organization.

1.3 Description of Tasks

1.3.1 Applied Linguistics

In the field of applied linguistics five separate tasks were pursued under this contract. These were:

Further expansion of the Russian Master Dictionary (RMD). This task consisted of correcting and modifying the information in the dictionary, as well as augmenting it with new entries.

Development of the Single-Pass Bidirectional translation system. This system, which made use of the RMD, was made operational during the course of the contract. It is capable of translating randomly selected Russian texts.

Development of the theory of microglossary organization. This scheme of storing microglossaries makes use of one common dictionary, and only words which are ambiguous are stored in separate glossaries.

Expansion of Multipass routines. The multipass routines developed during the previous year were further expanded to handle more complex grammatical structures.

Conversion of RMD to Multipass entries. This effort was directed toward establishing a program capable of converting automatically the entries in the RMD to the search input state entries and the final pass and backup entries.

1.3.2 Linguistic Research

The label "linguistic research" has been applied to a set of activities where deeper, as yet unformulated, aspects of grammar and notions of grammatical theory seem to impinge upon machine translation work. The activities of this nature that are currently considered to be of critical interest for machine translation within IBM Research are 1) housekeeping programs, 2) search strategies, 3) government information and 4) Russian grammatical studies.

It has become increasingly obvious that only complete control of sentence elements during automatic recognition procedures can guarantee success. A program designed to effect this kind of control is termed here a housekeeping program. Such a program should permit the manipulation and assignment to tree structures of the various syntactic elements of Russian sentences. It should also be consonant with the demonstrably most powerful theory of grammar, and it should be adaptable to a gradually expanding grammatical capability and to any search strategy. A housekeeping program designed to answer these demands has been studied and is being elaborated at IBM Research. A description of its features is included within the body of this report.

A recognition routine seems to be compounded out of a search strategy and a set of grammatical rules. Given the totality of grammatical rules for Russian, in particular, investigators at IBM Research are working to isolate the problem of a search strategy to determine the distinctive features of various search strategies, and to determine which of these strategies may be optimum or to create an optimum strategy.

The concept of government, especially within a highly inflected language like Russian, embraces countless grammatical facts that can contribute much to the improvement of Russian-English MT output. Even though government information has many deep implications within the structure of Russian grammar, it seems expedient to gather the data for simultaneous and/or subsequent study. For this purpose a very broad program for the mass compilation of government data has been outlined in this report. Further study and refinement of this program is envisioned before the mass compilation is initiated by a group such as the lexicographers at the Library of Congress.

To ensure the continuing improvement of the existing translation system, grammatical research in Russian has been intensified. A method of study is proposed which should lead to the rigorous formulation of sound grammatical rules for the Russian language. Some steps, at least, have been taken toward the development of a substantial set of constituent-structure rules for kernel sentences and a fair-sized set of transformational rules. This ever increasing body of data will represent a prime source of sound grammatical information for the expanding production system.

1.3.3 System Organization

At the completion of the Word Analyzer modification, the Mark II (AN/GSQ-16 language processor) had been equipped with the most powerful data processing capabilities considered applicable to natural languages. These capabilities have been extensively tested in the analysis of Russian grammar performed in the machine translation of Russian to English.

During this period of test, the new processing procedures were thoroughly evaluated in regard to effectiveness and efficiency. As part of this improvement program, the entire system organization has been reviewed and analyzed with particular attention being directed toward those areas where increased efficiency would be desired in a production model. This study has concentrated on the better utilization of the core memory (lexical buffer) which has become a central component of the system. Accordingly, a detailed study and logical design have been completed which exploit this component in the performance of rapid subtable searches entirely within the core memory. An estimate of anticipated production rates has also been prepared for multipass translation of Russian to English.

Table 2-2 Intermediate Passes - Sentence Analysis Statistics

Pass No.	Total No. of Entries, Average No. of Characters Per Entry	Total Table Size in Characters	No. of "Control" Entries, Avg. No. of Characters Per Entry	No. of Searches on "Control" Entries	No. of "Additional" Entries	No. of Searches on "Additional" Entries	No. of Main Dictionary Searches	No. of Core Table Searches
0	100 Entries, 20 Char. /Entry	2,000	100 Entries, 20 Char. /Entry	22	—	—	—	22
1	500, 35	17,500	70, 20	47	430	6	—	53
2	240, 40	9,600	240, 40	29	—	—	—	29
3	1060, 35	37,000	50, 25	25	1010	1	1	25
4	400, 35	14,000	400, 35	29	—	—	—	29
5	180, 35	6,300	180, 35	27	—	—	—	27
6	300, 40	12,000	300, 40	24	—	—	—	24
7	Undefined	33,300						32
8	700, 40	28,000	50, 25	22	650	1	—	23
9	3000, 45	135,000	500, 25	44	2500	2	2	44
10	360, 25	9,000	75, 20	31	285	2	—	33
11	Undefined	33,300						32
12	1600, 30	48,000	250, 20	33	1350	8	8	33
13	Undefined	33,300						32
14	Undefined	33,300						32
15	2700, 30	81,000	300, 20	40	2400	1	1	40

Note 1: Data for the four linguistically undefined passes are based on the average statistics of the twelve (12) defined passes.

Note 2: For a 20-word sentence, a total of 522 searches is made for passes 0 through 15. For the AN/GSQ-16 System the breakdown of searches shown in the last two columns is for an 8,000-word (36 bits/word) memory, with core table size restricted to 35,000 characters, providing for processing two sentences, of 150 words each, at a time. In this case there are 510 core table searches and 12 Photostore searches.

Note 3: With a 16,000-word memory, there would be 84,000 character addresses for core tables. In this case there would be 520 core table searches and 2 Photostore searches.

Note 4: Total storage requirement for passes 0 through 15: 533,000 characters.

Section 2: DISCUSSION

2.1. Applied Linguistics

2.1.1. Russian Master Dictionary (RMD)

The Russian Master Dictionary (RMD) now in use contains approximately 137,000 entries, each consisting of 1) an acquisition number, 2) an argument containing a sequence of symbols designed to provide a correct match for the Russian input data, 3) a function containing a translation, and, if necessary, information that indicates where the Russian ending begins, and 4) a confix, giving grammatical information about the Russian argument and English translation. If the argument is neither a verb, noun, adjective nor adverb, the confix field may be empty. Moreover, verbs that receive a nonpassive translation when they occur with a reflexive suffix are stored with a prefix in the argument field preceding the sequence of characters designed to match with the Russian input. The translation for the nonreflexive form of the verb is stored in the function field of an entry with an argument preceded by the prefix (Q8) RN, while the translation for the reflexive form of the verb is stored in the function field of a corresponding entry whose argument begins with a (Q8)RR. Verbs whose English translation is irregular have, in addition to the regular verb entries, special prefixed entries that contain in the function field the irregular translations for the forms requiring it. Entries of the last type have been added to the dictionary only very recently and will be discussed below. The following are typical RMD entries:

A.

ØØ9334	XOROW	GOOD	(R7) JU

B.

ØØ81Ø4	VECER	EVENING	(R1) WR

C.			
ø21259	GORI	(DP3) BURN	(R4) DC
D.			
øø7752	V#SILU#OPREDELEN14	BY#DEFINITION	(RM) NU
E.			
ø17ø48	MEJDU	BETWEEN	
F.			
ø169ø5	MEN6WE(,)#CEM	LESS#THAN	
G.			
ø41692	POVERXNOST6#KASANI4	TANGENT#SURFACE	
H.			
ø17914	(Q8)RN+POL6ZOVA	TREAT	(R4) LC
I.			
ø1938ø	(Q8)RR+POL6ZOVA	US	(R4) AE
J.			
173747	(R3)VBV+LOMIL	(DP3)BROKE	(R6)VBVXUXO
K.			
173749	(R3)VBD+LOMI	(DP3)BROKEN	(R6)VBDXUXO

(Here ø is a symbol standing for "zero," used to distinguish zeros from the letter O.)

The first three entries are typical adjective, noun, and verb entries ,respectively, while entry D although not an adverb in the classical sense always acts as an adverb, and is given an adverbial confix. Entry E is for a preposition and is thus not confixed. Similarly, entry F cannot be confixed as it acts neither as a verb, noun, adjective, nor adverb. Entry G, while acting as a noun cannot be considered as a noun, since it declines internally and the present translating system does not handle such a case. The next two entries are for the verb POL6ZOVAT6 which is translated "treat" when it occurs nonreflexively and "use" when it occurs reflexively. The last two entries are for the irregular past tense and past participle of "break."

In the last few months, many extensive changes beyond the usual additions, corrections, and deletions have been made as a result of dictionary proofreading and analysis of translation output.

Initially, lexicographers eliminated the most glaring excesses in assignments of multiple choices and deleted completely archaic and very colloquial entries. Approximately 16,000 RMD entries were affected. Such changes as the following tended to make the output much more readable.

TRI (PN)	(DP3) THREE/RUB	
TRI (PN)	(DP3) THREE	
SOROK (PN)	(DP5) FORTY/MAGPIES	
SOROK (PN)	(DP5) FORTY	
POCTI (PN)	(DP5) ALMOST/HONOR	
POCTI (PN)	(DP5) ALMOST	
POD (PN)	(DP3) HEARTH/UNDER	
POD (PN)	(DP3) UNDER	
VOZRASTU (PN)	(DP8) WILL#GROW/AGE	
VOZRASTU	(DPF) AGE	(R1) SR
VES6	VILLAGE/WHOLE	
VES6	WHOLE	

The next improvement to the dictionary involved approximately 25% of the entries. Originally, the machine lacked the ability to rematch on input material it had already matched on, consequently certain entries have to be entered as full forms or as longer stems, i.e., as complete words, or as forms containing part of an ending. For instance, GOVORI and GOVOR4 were included in the dictionary as longer stems to prevent the forms of the verb GOVORIT6 from matching on the entry for the noun GOVOR. But the verbal ending T is ambiguous, for it might come either from GOVORIT or GOVOR4T and the translation (SAYS/TALKS or SAY/TALK) can be resolved only by assigning GOVORI to a special inflexional class. Similarly, in the case of GOVORA, etc., if the full form had been entered to prevent conflict with the verb GOVORIT6 whose stem could then be entered as GOVOR, then, since the machine could not rematch on A, the information that GOVORA is in the genitive case would be lost and WUM GOVORA would be translated as "noise talking" rather than as "noise of talking." This deficiency prevented grammatical analysis whenever different Russian words (most often different parts of speech) had the same stem, as is often the case. To solve this problem, a new way to store this type of entry was developed. This involved adding in the beginning of the function field a so-called "DP" instruction which instructed the computer on how much of the matched information it should shift out. In almost every case the entry also needed to be reconfixed. After these two major changes, the output improved considerably.

For instance,

"At/during 'construction' molecules of fuel are used basically three/rub building material: carbon, hydrogen and oxygen,"

became

"At during 'construction' of molecule of fuel are used basically three building material: carbon, hydrogen and oxygen."

The next major change was the storage and confixing of verbs with irregular English translations. These entries were needed to take advantage of an unused capability of the Bidirectional Single-Pass Translation System, the technical details of which will be explained in Section 2.1.2.

Until the bidirectional translation system was developed, verb forms that received an irregular English translation had to be entered as full forms, thus losing the advantage of being able to go through the bidirectional verb analysis. Thus, "ON NE LOMAL STU L64" was translated "he not broke chairs" rather than "he did not break chairs," while "ON ZABYL KUWAT6" was translated "he forgot eat" rather than "he forgot to eat." Approximately 4000 entries were affected by this last dictionary change which also brought about a noticeable improvement in the translation output.

While the longer stems and full forms were being re-confixed, entries were being made for phrases (such as entry G above) which were set aside for use when the Bidirectional System could be re-written at which time the processing of such forms would be incorporated. These new entries indicate the end of the stem with a DP instruction and have a confix for the phrase to indicate that it is a phrase and also to give its part of speech. For instance the new entry corresponding to entry G is,

041692 POVERXNOST6#KASAN14 (DP10) TANGENT#SURFACE (R2) SUP

Here "(DP10)" indicates that POVERXNOST6 is the basic word in the phrase, its stem ending with the tenth character of the phrase. The confix (R2) SUP gives grammatical information about POVERXNOST and its English translation and indicates that the entry is for a phrase. There are over 5000 such entries ready to be added to the dictionary.

A minor change made in the dictionary a few months ago involved the deletion (by machine) of certain unnecessary and unused verbal stem entries (namely nonprefixed stem entries for the so-called R-verbs). Of course, this step brought about no alteration in the output.

It should be noted that there are now on hand for the RMD many additions that are currently in the final stages of processing and that soon will be included in the dictionary.

2.1.2. Bidirectional Analysis

2.1.2.1 Introduction During the past year the Bidirectional Single-Pass System with two-stem storage was developed and made fully operational. Later, both the linguistic and technical capacities of the system were enlarged to some extent, as will be explained below.

The Bidirectional Single-Pass translation program now is in the process of being completely rewritten to take advantage of the speed and logical uniformity possible with single-stem storage. In the new system each dictionary entry will be stored only once. Moreover, the linguistic capacities of the system will be significantly augmented, including, among other things, the skipping of adverbs and the use of microglossaries to resolve semantic ambiguities.

2.1.2.2 Current Linguistic Capabilities of the Bidirectional Single-Pass System. The Bidirectional Single-Pass translation system consists primarily of immediate environment analysis and is designed to eliminate the most common ambiguities in the Russian language and to smooth out the English translation. From the beginning the operational system has accomplished the following linguistic work:

Genitive and instrumental noun pairs are linked and OF or BY is inserted between the two elements, depending on whether the second noun is in the genitive or instrumental case, respectively. If the case of the second noun is ambiguous, the ambiguity is resolved in favor of the genitive and instrumental. If the ambiguity involves both the genitive and instrumental cases, OF/BY is inserted between the words.

In general, a noun is always translated as singular unless it is unambiguously plural, or unless the ambiguity is resolved by one of the other routines.

If, in linking noun pairs, the second noun is potentially genitive plural, it will be given a plural translation and OF will be inserted.

If a potentially plural noun follows a potentially plural adjective that agrees with it in case, with no punctuation marks intervening between the noun and the adjective, it is given a plural translation.

If a potentially plural noun is followed by a comma and then by an adjective agreeing with it in case and number, the noun is given a plural translation.

All adjectives following a noun and ending in OGO, EGO, EGOS4, YX, IX, IXS4 are construed to be genitive and have OF inserted in front of them; those ending in YM, IM, IMS4, OH, EH, EHS4, YMI, IMI, IMIS4 are treated as instrumental and have BY inserted in front of them; adjectives ending in O1E1E1S4 are considered to be ambiguous and are preceded by OF/BY.

All singular short-form adjectives following a singular nominative noun with which they agree in gender and number have IS inserted before them; plural short-form adjectives have ARE inserted when they follow a plural nominative noun.

No attempt is made to resolve ambiguous adverbs derived from adjectival stems, but clearly adverbial forms are translated as adverbs.

The adjective SAMY1, when followed by an adjective agreeing with it in case and gender, is translated as THE MOST; otherwise, its translation is ACTUAL.

Prepositions whose translation varies with the case they govern are linked with nouns or adjectives which follow them. When possible the proper meaning for the preposition is selected on the basis of this linkage. Some singular/plural ambiguity in nouns may also be resolved by this linkage.

Verbs go through a relatively complex analysis before being translated. First, they are classified on the basis of reflexivity. The so-called "R verbs," consisting of those which change meaning with the reflexive suffix S4 or S6, and those which occur only in the reflexive form, are not translated passively. The "non-R verbs." include all other verbs, which, when they are reflexive, are translated passively. It also includes those verbs which occur only in the non-reflexive form. (Both the reflexive and non-reflexive occurrences of

R-verbs are treated in the same way as the non-reflexive forms of non-R verbs. The discussion that follows gives the treatment of non-R verbs.)

In addition, all verbs are classified on the basis of aspect, so that there are three groups of R-verbs and three groups of non-R verbs: perfective, imperfective, and those that can be either perfective or imperfective, depending on the ending.

The present-tense form of perfective verbs is taken to be future and WILL is inserted before the English meaning. The third person singular of present tense imperfective verbs is translated by the "S form" of the English verb.

Their passive counterparts receive proper translation.

Most participles, with the exception of those whose translation differs significantly, are derived from the verb stems. They are formed by adding ED to the English verb stem for all participles other than the non-reflexive active participles which are formed by attaching ING to the English stem.

Gerunds are translated by adding ING to the English stem. For reflexive gerund of non-R verbs, BEING is inserted and ED is added to the stem.

The infinitival TO is not normally inserted, except when the infinitive is preceded by a noun or another verb (all forms but the participles). If the infinitive has S4 at the end, and if it belongs to the non-R class of verbs, TO BE is inserted and ED is added to the stem.

All other verbal forms are translated in the usual way, the appropriate passive translation rendering the reflexive suffix for the non-R class of verbs.

When a verb is immediately preceded by NE, with no punctuation marks separating the two words, the translation of the verb is modified by the correct English form. All the rules discussed are followed, except that the inserted auxiliary verbs are the negative English forms.

When Russian BY directly follows a past tense verb, WOULD is inserted, and the verb is translated in the present-tense form. If NE precedes a past tense verb which is followed by BY, WOULD NOT is inserted before the present tense form of the English translation. When, in both of the above cases the verb is a reflexive non-R verb, WOULD BE or WOULD NOT BE is inserted before the "ED" form of the English verb.

When the verb STAT6 is followed by an infinitive, its semantic translation is BEGIN. The meaning is then properly inflected to render the tense, person, etc., of the particular form of STAT6. When STAT6 is not followed by an infinitive, its semantic translation is STAND/STOP.

All present tense forms of BYT6 (i.e., BUDU, BUDEW6, etc.) are translated as WILL when followed by an infinitive. In this case TO is not inserted before the infinitive. When anything else follows these forms of BYT6, their translation is WILL BE.

When NE precedes STAT6 or BUDU, the translation becomes the correct counterpart of the above-mentioned forms.

Two important linguistic capabilities were added to the translating system after it was already operational.

The first involved word derivation. When a word is not in the dictionary but is derived by a common rule from another word that is in the dictionary, the program can recognize this fact, generate the correct translation, and then utilize all the grammatical information latent in the word's grammatical ending.

The following derivations are made:

Russian Suffix	English Translation for Suffix
1. Noun to Noun	
IST	IST
IT	ITE
IZM	ISM
K	
W	
ISTK	ISTICS
IK	ICS
Q	
2. Noun to Adjective	
ICN, ICEN	IC
CN, CEN	IC
ICESK	IC
CESK	IC
VIDN, VIDEN	-LIKE
OBRAZN, OBRAZEN	-LIKE
PODOBN, PODOBEN	-LIKE
NOSN, NOSEN	-BEARING

Russian Suffix	English Translation for Suffix
N	
4N	
AN	
AL6N	AL
ONAL6N	ONAL
IVN, IVEN	IVE
EN, OCN, OCEN	
SK	
OV, OVSK	
EV, EVSK	
3. Adjective to Noun	
OST	NESS
4. Adjective to Adjective	
OVAT	ISH
EVAT	ISH
E1W	EST
A1W	EST
N	
5. Verb to Noun	
ANI	ING
ENI	ING
AQI	ATION
4QI	ATION
6. Verb to Verb	
L4	
OVYVA	
VA	
IVA	
YVA	

These derivations have been very useful with unanticipated and newly coined words but also have the important function of limiting the size of the dictionary by eliminating the need for entering many semantically related words in the dictionary.

The following are examples of typical derivations:

The stem NAPISAN (WRITTEN) can be used for the masculine short form participle NAPISAN (WRITTEN) and for the masculine participle NAPISANNY1, since N is a suffix in the adjective-adjective derivation category. The noun PISANIE (WRITING) can be derived from the verb stem PIS for the verb PISAT6 (WRITE). The superlative adjective XOLODNE1WI1 (COLDEST) can be derived from the adjective stem XOLODN (COLD). The noun VESELOST6 (GAYNESS) can be derived from the adjective stem VESEL (GAY). The nouns LENINIZM (LENINISM) and STUDENTKA (STUDENT) can be derived from the noun stems LENIN (LENIN) and STUDENT (STUDENT) The adjectives ARTISTICESKI1 (ARTISTIC) and TRUBOVIDNY1 (PIPE-LIKE) can be derived from the stems for the nouns ARTIST (ARTIST) and TRUBA (PIPE).

The second important linguistic improvement in the operational Bidirectional system consisted in the utilization of its built-in capacity to handle irregular English verbs. Until recently, Russian verbs that had irregular English translations were entered as full forms or longer stems and thus were not able to take advantage of the Bidirectional analysis. Thus, POWEL KUWAT6 was translated WENT EAT rather than WENT TO EAT, etc. Moreover, since almost every Russian verb that had an irregular English past tense and past participle - the passive and negative forms being the most numerous here - would have required over 30 separate entries to insure a correct translation of all the forms, usually all of these forms were not entered, semi-correct translations being preferred to an over-bulky dictionary. In fact entries were usually limited to full forms or longer stems for the Russian past tense and passive participles. Even then, the negative past tense usually was not entered and came out incorrectly, NE POWEL being translated NOT WENT rather than DID NOT GO. By adding three new tables to the dictionary, (a table for the irregular past participles of non-R verbs that can occur reflexively, a table for all the irregular forms of non-R and non-reflexive R-verbs, and a table for all the irregular forms of reflexive R-verbs), and making new entries for the Russian verbs with irregular English translations, it was made possible for all the forms of Russian verbs with irregular English translations to be translated correctly. These verbs now go through the regular Bidirectional verb analysis, as will be explained in detail in the following section.

2.1.2.3. Programming Aspects of the Bidirectional Single-Pass Translation System Currently in Operation. The format of the dictionary used in the Bidirectional analysis differs from the format of the RMD but the conversion of the entries is performed with the use of a computer. Full forms remain full forms but other kinds of RMD entries correspond to two or more entries in the Bidirectional dictionary analysis, as will be shown.

The original Bidirectional program has undergone minor changes (word derivation has been added, for one) since it was first operational. However, one very significant improvement that has been incorporated has greatly increased the speed of translation. To understand the reasoning involved here, one must first know something about the machine logic and about the storage of the Bidirectional dictionary and control entries.

All the lexical information and control information used in the Bidirectional System is stored serially along concentric tracks on a rotating glass disc. When a word is being looked up, the disc is read by a cathode-ray tube light source which steps across tracks, sampling a small portion of each, until a comparator indicates that it has gone too far. The beam is then brought to rest and the disk rotation allows the reading of every entry on that track until a proper match is made.

Now, the original method of storing Russian stems in the Bidirectional dictionary included a table of all the stems (approximately 150,000) preceded by the prefix (R3)ØØØ. The logic of the machine allows any character to be considered a correct match with a zero, ((R3)VNE+BIT6) thus being considered by the machine a correct match with((R3)ØØØ + BIT6), the three zeros following the (R3) being used to "mask over" confix information about the previous word. However, the logic of the machine also requires the search for a word in this table to start at the beginning of the table. This occurs because of the zeros which cause an entry search above them, and there are good reasons for this logical requirement. But the long entry search causes the lookup time per word to be a matter of seconds, while a regular track search through a table of the same length requires only a few milliseconds. A programming method has been devised which will avoid storing dictionary entries in tables preceded with prefixes containing zeros. This method will be used in the new (not yet operational) Bidirectional System with single-stem storage and will be explained along with the rest of the program in Section 3.1.2. A temporary and minor hardware change was made to overcome this problem until the new Bidirectional program would become operational.

A general discussion of the current programming aspects of the Bidirectional Single-Pass system follows:

1. Nouns

Noun stems are now stored in the dictionary in two forms:

STEM	TRANSLATION	$(R3)G_1G_2G_3$
(RZ)ØØØ + STEM	TRANSLATION	$(R3),,,L_SG_1G_2G_3$

where (R3) is a confix used in noun analysis, $G_1G_2G_3$ gives grammatical information about the Russian stem and the English translation, L_S is the length of the noun stem, Ø (zero) is the "suppress matching" instruction, "," is the "copy not" instruction, and + serves to separate the prefix from the stem.

Each noun ending appears in the dictionary in the form:

$(R3)G_1G_2$Ø + END(PN) (DPL_E) (QØ)No n ,

where Ø serves to match on G_3 in the noun confix and "," leaves G_3 unchanged in the confix region of memory, (DPL_E) shifts out the ending (leaving the pointer at (PN) which is the first byte of each punctuation code), No stands for number (singular or plural), and n indicates whether the case of the Russian ending is potentially nominative and, if it is, also indicates the gender and number of the noun.

The confix beginning with (QØ) is used to read out the proper

English suffix to be attached to the translation, and is designed to match on an entry of the following type:

$(Q\emptyset)$ No$\emptyset G_3$ + SUFFIX (R3)N, C .

The confix stuffed, (R3) N n C indicates that the word just translated was a noun (N), at the same time indicating if it is nominative, etc.

In addition to being stored as are all other noun endings, all ambiguously plural noun endings entered in the dictionary are also followed by a comma and a space:

(R3)$G_1 G_2 \emptyset$ + END(,)# (R3)XCa, ;

here Ca stands for case and X is a confix which identifies this confix as being used in connection with the analysis of potentially plural nouns (followed by a comma) that will be given later.

Noun-adjective and noun-noun derivational suffixes are now entered in the dictionary in the following forms:

(R3)$G_1 \emptyset\emptyset$ + $SUFF_N$ $(Q\emptyset)G_1{}'G_2{}'$, tZN

(R3)$G_1 \emptyset\emptyset$ + $SUFF_A$ $(Q\emptyset)G_2{}'$A, tZN

where $G_1{}'$ and $G_2{}'$ give the grammatical information about the new nouns and adjectives formed, and t specifies the translation to be given for the suffix. The confix directs the next match to an entry of the form:

$(Q\emptyset)\emptyset\emptyset G_3$ $\emptyset$ Zn + ENG. $SUFF_{G3}$ $(Q\emptyset)$, ,R, ZN

which completes the English translation for the basic stem - CIT (CITY-CITIES) becomes CITY - and the next match will be on one of the following entries:

$(Q\emptyset)\emptyset\emptyset\emptyset$ t_N ZN + ENG. $SUFF_{t_N}$ (R3), ,$G_3{}'$

(QØ) ØØØ t_A ZN + ENG. $SUFF_{t_A}$ (R7), G_3'

whose confixes will direct the next matches to noun and adjective-ending tables, respectively.

If a noun directly follows another noun, its lookup will be modified by the "noun-preceding confix" mentioned above. This confix (as all other confixes carrying over information from a preceding word) matches on an entry that is designed to prevent the next search from being a "long entry search," and that changes the initial (R3) to (RZ). Consequently, the noun will not match on the simple stem entry, but on the entry.

(RZ)ØØØ + STEM TRANSLATION (R3),,, $L_N G_1 G_2 G_3$

After this match is made and the translation is read out, the confix will match on an entry which erases the translation and changes the confix so that the next match will be on one of the following types of entries:

(R3)NBØØG_1G_2 + END_G(PN)	(DPØ)OF#	(R3)BB,,
(R3)NBØØG_1G_2 + END_I(PN)	(DPØ)BY#	(R3)BB,,
(R3)NBØØG_1G_2 + $END_{G/I}$(PN)	(DPØ)OF/BY#	(R3)BB,,
(R3)NBØØØØ +		(R3)BB,,
(R3)NBØØG_1G_2 + $SUFF_N$		(R3)ZN,,$G_1'G_2'L_E$
(R3)NBØØG_1G_2 + $SUFF_A$		(R3)ZA,,G_1'X L_E

Here END_G, END_I, $END_{G/I}$ are the genitive, instrumental, and genitive and instrumental endings for class G_1G_2, and $SUFF_N$ and $SUFF_A$ are nominal and adjectival suffixes which, when attached to a given stem, form nouns of adjectives of classes $G_1'G_2'$ and G_1', respectively. L_E stands for the length of these suffixes.

If a match is made on one of the first four entries the next match will be on one of the so-called "return entries," which shift the attention of the computer to the beginning of the second noun's stem. If a match is found with one of the derivational entries, the next match will be made on a "length tag addition" entry of the form:

$(R3)Z\emptyset\emptyset L_N\emptyset\emptyset L_E +$ $(R3)Q,,L_{N+E}$,,

which provides the new length tag and then channels the search to one of the first four entries shown above for nouns, and to a corresponding set of entries for adjectives, that will be explained later.

If the ending of a noun preceded by another noun is potentially plural genitive, the path taken differs somewhat from the one outlined above, and after matching on return entries, a match is found with an entry that leaves the confix (R3) APG. This confix insures a plural translation of a potentially genitive plural noun followed by it and is of the same type as that stuffed by all plural adjectives.

All plural or potentially plural adjectives, when followed by a space, stuff a confix of the form (R3) APCa. If a match is made on a noun, the translation will be read out, and the next match, designed to test if the noun and adjective should be linked together, is on one of these entries:

$(R3)APN\emptyset G_1G_2\emptyset + END_N(PN)$ (DPL_E) (R3)PLURLP,

$(R3)APCa\emptyset G_1G_2\emptyset + END_{Ca}$ (PN) (DPL_E) (R3)PLURLA,

$(R3)AP\emptyset\emptyset G_1G_2\emptyset +$ $(R3)G_1G_2SING$,

where END_N is a nominative ending, Ca stands for case, (DPL_E) shifts out the ending matched on, D and A stand for nominative and non-nominative plural, respectively, and PLURL and SING indicate whether the suffix to be attached to the translation should be plural or singular.

As was mentioned above, all ambiguously plural noun endings entered are followed by a comma in order to link these nouns with the adjective, agreeing in case and number, that follows. If the following word is not an adjective, a singular translation is given to the noun; if a match is found with an adjectival stem, the next match made will be either with

(R8)XCaØØG_1 + END_{Ca} (PN)	(DPØ)		(R8)XP,,BA
or			
(R8)XØØØØ +			(R8)XS,,BA

depending on whether or not the adjective agrees in case with the noun. If it does, a match will be found with the first entry that assures a plural translation for the noun ending and a shift back to the beginning of the adjectival stem. If the adjective does not agree in case with the noun, a match will be found with the second entry. This results in a singular translation for the noun and a shift back to the adjectival stem.

Noun-adjective, adjective-adjective, and verb-participle derivation are provided for with a procedure similar to that mentioned above, a match being made with a "length-tag addition" entry, which then refers the search to the entries that investigate the adjectival endings for case agreement with the noun.

2. Adjectives

All adjectival stems also are stored in the dictionary twice:

STEM	TRANSLATION	(R7)G_1G_3
(RZ)ØØØ + STEM		(R7),,,L_SG_1

where, as before, G_1 and G_3 carry grammatical information about the stem and translation, and L_S gives the length of the stem.

All adjectival endings are stored in the form:

(R7)G_1∅ + END(PN)   (DPL$_E$)   (A∅)F,

where (DPL$_E$) again shifts out the ending (END) and F stands for adjectival form, P for positive, C for comparative, or A for adverbial. The confix (Q∅)FG_3 matches on

(Q∅)FG_3 +   SUFFIX

In addition, all potentially plural endings are stored as

(R7)G_1∅ + END#   (Q∅)X, Ca

where X is a constant and Ca stands for case. This confix is designed to match on

(Q∅)XG_3∅ +   POS. SUFF.#   (R3)AP,

which then modifies the successive lookup.

Adjective-noun and adjective-adjective derivation is provided here in a way analogous to the method used with nouns.

If a noun precedes an adjective, the adjective will be looked up in the confixed form.

If the adjectival ending is genitive or instrumental, the next match will be found with one of the entries:

(R8)N∅∅∅G_1 + END$_G$(PN)   (DP∅)OF#   (R3)BB,,
(R8)N∅∅∅G_1 +END$_I$ (PN)   (DP∅)BY#   (R3)BB,,
(R8)N∅∅∅G_1 + END$_{G/I}$(PN)   (DP∅)OF/BY#   (R3)BB,,

which then insures a shift back to the adjectival stem. If the ending is neither genitive nor instrumental (nor a derivational suffix, a procedure being provided here similar to that for nouns), there is a shift back to the beginning of the adjective stem, provided a match is not found with the entries discussed in the next paragraph.

In order to insert IS or ARE in front of short-form adjectives the short-form endings are entered in the form:

(R8)Nn∅∅G_1 + $END_{n_{(S)}}$	(PN)	(DP∅)IS#	(R3)BB,,
(R8)Nn∅∅G_1 + $END_{n_{(P)}}$	(PN)	(DP∅)ARE#	(R3)BB,,

after which there is a shift back to the adjectival stem.

The adjective SAMY1 is entered in the dictionary in two ways: as a stem SAM, and in all the full inflected forms followed by a space:

SAM	ACTUAL	(R7)G_1G_3
(R3)∅∅∅ + SAM		(R8),,,L_SG_1
and		
SAMY1#		(R3)SMY_1
SAMA4#		(R3)SMY_2
SAMOE#		(R3)SMY_3

and so on.

Here SM stands for SAMY1, and Y_1 demotes the unique morphological form of SAMY1.

If SAMY1 is followed by a space but not by an adjective, ACTUAL# is given as a translation.

If an adjective does follow SAMY1, it will be looked up in the prefixed form, and if it doesn't agree with SAMY1 in case, number, and gender, then the machine will be shifted back to the beginning of the adjectival stem and ACTUAL# will be given as a translation.

If the adjective does agree with SAMY1, a match will be found with,

(R8)SMYØG_1 + END_Y(PN) (DPØ)THE#MOST# (R8)SBA,

which then shifts back to the beginning of the adjective with the aid of the confix (R8)SBA,.

Noun or verb stems following a form of SAMY1 will cause a translation of ACTUAL# unless they are followed by adjective derivational suffixes in which case the adjective formed will be checked for agreement with SAMY1 with the same entries used above.

3. Prepositions

Prepositions whose meaning changes with case are stored in the dictionary as,

PREP(PN) (DPLP)Ta/Tp

and

PREP# (R3)PpA

where PREP stands for preposition, LP for the length of the preposition, and Ta and Tp for the two meanings associated with the cases a and b; in the confix the P and A are constants, and p describes the preposition.

If a verb follows the preposition, the ending is investigated for participial suffixes. If one is found, the search is referred to the adjectival endings for investigation of case agreement. If no participial suffix is found, there is a shift back to the beginning of the verb and the ambiguous translation is given for the preposition.

If a noun follows a preposition, it will be looked up in the confixed form and the next match will be on,

(R3)PØAØØØØ + (TF) (R3)P,E,,,

where the function of (TF) is to erase the translation read out on the previous match.

The following match will be made with,

(R3)PpEØØØ + (R3)PpF, , , ab

where a and b denote the cases taken by the preposition p, and F is a constant.

This leads to a match with one of the entries:

(R3)PØFØG_1G_2ab + END_a(PN)	(DPØ)	(R3)P, G, aNo
(R3)PØFØG_1G_2ab + END_b(PN)	(DPØ)	(R3)P, G, bNo
(R3)PØFØØØØØ +		(R3)P, H, UN
(R3)PØFØØØØØ + $SUFF_A$		(R3)P, C, G_1L_S
(R3)PØFØØØØØ + $SUFF_N$		(R3)P, J, $G_1G_2L_{NS}$

where No stands for number (S or P,) UN is a constant meaning unsolved," $SUFF_A$ and $SUFF_N$ are adjectival and noun suffixes, respectively.

If a match is made on one of the last two entries, the next match will be made with one of the "length-tag addition" entries which then refer the search either to noun entries that investigate the ending for agreement or to adjectival entries of the same type. If a match is made with the third entry, the resultant confix will match on a "return entry," assuring the correct ambiguous translation for the preposition. And, if a match is made with either of the entries that determine the case by the preposition, the machine will shift back until a match is found with one of the entries:

(R3)PpIAaS + PREP# Ta#

(R3)PpIAbS + PREP# Tb#

(R3)PpIAaP + PREP# Ta# (R3)APCa

(R3)PpIAbP + PREP# Tb# (R3)APCa

where (R3) APCa is a confix assuming a plural translation of an ambiguously plural noun whose number is resolved by the government of the preposition.

If an adjective follows a preposition,it will be looked up in the confixed form and the next match will be made with,

(R8)PpAØØ + (R8)PpB, , ab

where a and b again stand for the two cases that the preposition p can take. If the adjective does not agree in case with the preposition, there is a shift back to the beginning of the adjective and the ambiguous translation is given for the preposition. If the adjective does agree in case with the preposition, a match will be found on one of the two entries:

(R8)PØBØG_1ab + END_a(PN) (DPØ) (R3)P, G, aS

(R8)PØBØG_1ab + END_b(PN) (DPØ) (R3)P, G, bS

The search then continues in the same way as for nouns. Finally derivational entries are provided for nouns or adjectives derived from adjectival stems,

(R8)PØBØØØØ + $SUFF_N$ (R3)P, J, G_1'G_2'L_{NS}

(R8)PØBØØØØ + $SUFF_A$ (R8)P, G, $G_1$$L_{AS}$

These entries act the same way as the derivational entries do for derivation from nouns.

4. Verbs

All R-verbs are stored in the dictionary in three forms:

STEM		(R6)VRaL_S
(RZ)ØØØ + STEM	N. TRANSLATION	(R6),,,L_SRaG_3
(R3)VRØ + STEM	R. TRANSLATION	(R6)VB,ERaG_3

where N and R translations are the English stems for the non-reflexive and reflexive occurrences of the verb, respectively; R indicates that the verb belongs to the class of R verbs; "a" stands for aspect and can be either P,I, or O,B; and E is a constant: the other symbols are as before.

Non-R verbs are stored twice:

STEM		(R6)VUaL
(RZ)ØØØ + STEM	TRANSLATION	(R6),,,L_SUaG_3

where the translation is in the English stem form, U indicates that the verb belongs to the non-R class, and all other symbols have the same meaning as for R verbs.

Verb endings, besides being used in the verb analysis which will be discussed in detail later, are also stored in a form which serves to read out their meaning as verb suffixes:

(R6)VBVØØØØ + END(PN)	(DPL_E)	(QØ)VgTERM,
(R6)VBVØØIØ + END_S(PN)	(DPL_E)	(QØ)VSTERM,
(R6)VBVØØOØ + END_S(PN)	(DPL_E)	(QØ)VSTERM,
(R6)VBVØØØØ + $SUFF_{PT}$		(QØ)G_1'gPART,
(R6)VBVØØPØ + $SUFF_{IMP}$		(R6) VBV,,I,
(R6)VBVØØØØ + $SUFF_N$		(QØ)VgNOUN,

where VBV is a constant distinguishing this class of entries, I, P, ∅ give aspect, END_S is a third person singular ending $SUFF_{PT}$, $SUFF_{IMP}$ and $SUFF_N$ are participial, imperfective, and nominal suffixes, respectively; g describes the morphological form of the English suffix to be attached to the translation, G_1' stands for the grammatical class of the participle, TERM, PART, and NOUN stand for "terminal," "participle," and "noun," respectively, and are used to distinguish the confixes. The resultant confix starting with (Q∅) is referred to entries which provide the English suffix:

(Q∅)VgTERM G_3 +	SUFF	(R3)VVV
(Q∅)VgPARTG$_3$ +	SUFF	(R7), G_3'
(Q∅)VgNOUNG$_3$ +	SUFF	(R3)G_1'G_2'G_3'

Here the suffixes are a function of g and G_3, (R3)VVV is a confix which conveys the information that the translated word was a verb, and (R7)G_1'G_3' and (R3)G_1'G_2'G_3' refer the search to the adjectival and noun endings, respectively.

We will first discuss the translation of <u>perfective R-verbs</u>. The confix stuffed by a perfective R verb is of the form (R6)VRPL$_S$. It appears in the dictionary as follows, alone, or with different entries which determine the translation of the verb:

(R6)VRP∅ + END_{Pr}(PN)	(DP∅)WILL#	(R6)BBE,
(R6)VRP∅ + $END_{Pr}SUFF_R$(PN)	(DP∅)WILL#	(R6)BRE,
(R6)VRO∅ + END_{PrPer}(PN)	(DP∅)WILL#	(R6)BBE,
(R6)VRO∅ + $End_{PrPer}SUFF_R$(PN)	(DP∅)WILL#	(R6)BRE,

(R6)VRØØ + (R6)BBV,

(R6)VRØØ + $Ø_E$SUFF$_R$(PN) (DPØ) (R6)BRV,

(R6)VRØØ + $Ø_E$(PN) (DPØ) (R6)BBV,

where END$_{Pr}$ stands for "present tense form ending," END$_{PrPer}$ stands for "present perfective ending," SUFF$_R$ is the reflexive Russian suffix; $Ø_E$ stands for a series of zeros necessary to skip over all the characters of a verb ending, E indicates that the suffix to be attached should be in the "E" form, regardless of the Russian ending, V indicates that the English suffix will be determined by the ending, and the B and R in the third position of the confix denote that the meaning to be taken is the non-reflexive and reflexive, respectively.

The confixes resulting from matches on the above entries match on a set of "return entries" which shift back to the beginning of the verb stem and change the second position B to V. The next match will be on the stem prefixed by either (RZ)ØØØ + or (R3)VRØ +, depending on whether or not the ending was reflexive. After this match the verb endings are looked up again, matching on either the entries first discussed above, or on

(R6)VBEØØØG_3 + "E"SUFFIX (R6)(QØ)VEØ

which then leads to a set of entries that shift out the ending up to the next punctuation and stuff a (R3)VVV confix.

Imperfective R-verbs stuff a confix of the form (R6)VRIL$_S$. This confix is used to determine whether or not the verb is reflexive.

This determination is accomplished by the entries already mentioned above. The lookup following a match made on those entries is the same as that for perfective R-verbs.

Perfective non-R verbs stuff the confix (R6)VUPL$_S$, which matches on one of the following entries:

(R6)VUØØ+		(R6)BBV,
(R6)VUPØ + END$_{Pr}$(PN)	(DPØ)WILL#	(R6)BBE,
(R6)VUOØ + END$_{PrPer}$(PN)	(DPØ)WILL#	(R6)BBE,
(R6)VUPØ + END$_{Pr}$SUFF$_R$(PN)	(DPØ)WILL#BE#	(R6)BBD,
(R6)VUOØ + END$_{PrPer}$SUFF$_R$(PN)	(DPØ)WILL#BE#	(R6)BBD,
(R6)VUØØ + END$_{PaS}$SUFF$_R$(PN)	(DPØ)WAS#	(R6)BBD,
(R6)VUØØ + END$_{PaPl}$SUFF$_R$(PN)	(DPØ)WERE#	(R6)BBD,
(R6)VUØØ + END$_I$SUFF$_R$(PN)	(DPØ)BE	(R6)BBD,
(R6)VUØØ + END$_{Pt}$SUFF$_R$(PN)	(DPØ)	(R6)BBD,
(R6)VUØØ + END$_G$SUFF(PN)	(DPØ)BEING	(R6)BBD,

where END$_{PaS}$ and END$_{PaPl}$ stand for past tense singular and past tense plural endings respectively, END$_I$ for imperative and infinitive endings, respectively, and END$_{Pt}$ for gerundive endings; the other symbols are as before.

After this match is made, the search proceeds in the same fashion as shown before. Confixes with D in the fourth position match eventually on entries of the type,

(R6)VBDØØØG$_3$ +.	"ED"SUFFIX	(R6)QØ)VEØ

which again leads to a set of entries that shifts out the endings and that stuffs an (R3)VVV confix for all entries but participles.

The imperfective non-R verb confix (R6)VUILS is used in exactly the same way as that of the perfective non-R verbs, with the exception that the present tense ending results in different meanings and translations. Consequently, the entries for perfective non-R verbs shown above, which have ∅ in the fourth position, are used by both the perfective and imperfective non-R verbs and only the following entries are added, to be used exclusively by the imperfective verbs:

(R6)VUI∅ + $END_{PrI}SUFF_R$(PN)	(DP∅)AM#	(R6)BBD,
(R6)VUO∅ + $END_{Pr1I}SUFF_R$(PN)	(DP∅)AM#	(R6)BBD,
(R6)VUI∅ + $END_{Pr2}SUFF_R$(PN)	(DP∅)ARE#	(R6)BBD,
(R6)VUO∅ + $END_{Pr2I}SUFF_R$(PN)	(DP∅)ARE#	(R6)BBD,
(R6)VUI∅ + $END_{Pr3}SUFF_R$(PN)	(DP∅)IS#	(R6)BBD,
(R6)VUO∅ + $END_{Pr3I}SUFF_R$(PN)	(DP∅)IS#	(R6)BBD,
(R6)VUI∅ + $END_{PrP1}SUFF_R$(PN)	(DP∅)ARE#	(R6)BBD,
(R6)VUO∅ + $END_{PrP1I}SUFF_R$(PN)	(DP∅)ARE#	(R6)BBD,

where the numbers indicate the person of the verb ending and the other symbols are as before. The matches proceed as before.

The Russian particle NE is stored in the dictionary in two forms:

NE(PN)	(DP2)NOT	
NE#		(R3)VNE

If the word following NE# is a noun, adjective, or full form, NOT# is given as a translation and the confix is dropped.

For verbs preceded by NE# the analysis of endings is done in essentially the same way as shown in the preceding paragraphs. The confix under which the endings are stored is (R6)VNFØØØ +, the sixth and seventh position Ø's being sometimes replaced by R and U, or I, P, O, respectively, and depending on the type of ending the translation given in each case is the negative counterpart of those shown above. After the ending is investigated, the search proceeds in the same fashion as for verbs not preceded by NE.

To translate past tense R verbs followed by the Russian BY, the following entries are stored in the dictionary:

(R6)VRØØ + END_{Pa}#BY(PN)	(DPØ)WOULD#	(R6)BBE,
(R6)VRØØ + $END_{Pa}SUFF_{R}$#BY(PN)	(DPØ)WOULD#	(R6)BRE,

In addition to these entries, an entry is provided for shifting out BY when it follows a verb.

Further, another set of entries is provided to translate past tense R verbs preceded by NE and followed by BY:

(R6)VNFØØØ + END_{Pa}#BY(PN)	(DPØ)WOULD#NOT	(R6)BBE,
(R6)VNFØRØ + $END_{Pa}SUFF_{R}$#BY(PN)	(DPØ)WOULD#NOT	(R6)BRE

To translate past tense non-R verbs followed by BY, the above entries which do not have an R can be used in addition to the following entries:

(R6)VUØØ + END_{Pa}#BY(PN)	(DPØ)WOULD#	(R6)BBE,
(R6)VUØØ + END_{Pa}#BY(PN)	(DPØ)WOULD#BE	(R6)BBD,
(R6)VNFØUØ + $END_{Pa}SUFF_{R}$#BY(PN)	(DPØ)WOULD#NOT#BE	(R6)BBD,

Irregular English verbs share the entries listed above for verbs with regular English translations.

Non-R verbs with irregular English translations have (RZ)ØØØ entries (with the "E" translation) as do all other verbs. In addition, all their forms with irregular English translation are stored in entries of the type:

(R3)VBV + FORM	(DPn)IRR TRANSLATION	(R6)VBVXUXO

If they occur with a reflexive suffix, their irregular past participles are also stored in entries of the form,

(R3)VBD + STEM	PAST PART.	(R6)VBDXUXO

Here the XUX in the confix is a constant signifying that these are entries for non-R verb with an irregular English translation.

When the verb is first matched on,it stores the confix

(R6)VUIL$_S$, (R6)VUPL$_S$ or (R6)VUOL$_S$, as do other verbs. The ending investigation proceeds as before for word insertion with the resulting confix of (R3)VBE("E" translation of verb required) (R3)VBD (past participle required; match was a reflexive non-R verb), or (R3)VBV(ending must be investigated to determine the proper translation for the verb form). The next match is to get a translation for the verb form (which will be placed after the word inserted if there was one). If the confix is (R3)VBE, the match on the verb will be made in the (RZ)ØØØ table where the required "E" translation is stored and the processing will continue as for regular verbs. If the verb form is reflexive, the match will be in the (R3)VBD table if the verb's past participle was irregular. The confix (R6)VBDXUXO will share the entries

(R6)VBDØØØG$_3$ +   "ED"Suffix   (R6)(QØ)VEØ

used for regular verbs but the last character, O, of the confix will insure a vacuous "ED" suffix. The next matches will continue as for regular verbs. If the confix was (R3)VBV, the verb form will match in the (R3)VBV table if its translation is irregular. The confix (R6)VBVXUXO will direct the next match to the regular verbal endings table after which the matches will continue as for regular verbs. The last character, O, of the confix, as G$_3$, will insure a vacuous English suffix for the translation.

R-verbs with irregular English translations have a stem entry, an (RZ)ØØØ entry, and a (R3)VRØ entry, as do regular verbs. In addition, all the non-reflexive forms with irregular English translations have an entry:

(R3)VBV + FORM   (DPn)IRR. TRANS.   (R6)VBVXRXO, while all the reflexive forms with irregular English translations have an entry:

(R3)VRV + FORM (DPn)IRR. TRANS. (R6)VBVXRXO.

(Here the XRX in the confixes is a constant signifying that these entries are for an R-verb with an irregular English translation.)

When the verb is first matched on, it stores the confix (R6)$VRIL_S$, (R6)$VRPL_S$, or (R6)$VROL_S$, as do other verbs. The ending investigation for word insertion proceeds as before, with the resulting confix of (R3)VBE ("E" translation of non-reflexive form required), (R3)VRE ("E" translation of reflexive form required), (R3)VBV (ending must be investigated to determine the proper form of the non-reflexive translation to be used), or (R3)VRV (ending must be investigated to determine the proper form of the reflexive translation to be used).

The next match is to get the proper translation for the verb form (to be placed after the word inserted, if there was one). If the confix is (R3)VBE, the match on the verb will be made in the (RZ)ØØØ table where the required "E" form for the non-reflexive translation is stored, and the processing then continues as for regular verbs. If the confix is (R3)VRE, the match on the verb will be made in the (R3)VRØ table where the required "E" translation for the reflexive form is stored, and the matches that follow will be as for regular verbs. If the confix is (R3)VBV, the processing is as for non-R verbs with irregular English translations, while if the confix is (R3)VRV, the verb form will match in the (R3)VRV table if its reflexive translation is irregular. The confix, (R6)VBVXRXO, will direct the next matches to the same entries as the confix (R6) VBVXUXO for non-R verbs.

Since the verbs with irregular English verbs share all the endings tables with the regular verbs, their treatment when preceded by NE or followed by BY, etc., is the same as for the regular verbs.

As was already mentioned all participles are always treated as adjectives. Most of them, however, are derived from verb stems, and require "length-tag addition" entries in most analyses, after which the match can be directed to the appropriate adjective endings table.

If a noun or a verb precedes another verb, the verb will again be looked up in the confixed form. The translation is then erased, and the ending is analyzed to determine if it is infinitival. If it is, TO# is inserted, and there is a shift back to the beginning of the stem. If it is not, nothing is inserted before the shift is made.

STAT6, in addition to being entered as a stem, has all its full forms entered in the dictionary as follows:

STAT6#	(R3)STf_1
NE#STAT6#	(R3)STf_2

etc., where ST is a constant standing for STAT6, and f_i denotes the English grammatical form by which the particular form of STAT6 should be translated. If the word following STAT6 is a verb, the translation is erased, and the ending is checked. If it is infinitival, F is written in the seventh position of the confix, while if it is not, an N is stored in that place. Then a series of entries shift back to the beginning of the verb stem, after which our match is found on an entry of the type:

(R6)STf_iBIFB++	BEGIN#TOf_i
(R6)STf_iBINB++	STAND/STOP#f_i

If the word following STAT6 is a noun or adjective, the procedure is roughly the same as that for a verb which is not an infinitive, i.e., N is stored in the seventh position of the confix, leading to entries which translate STAT6 as STAND/STOP.

All present tense forms of BYT6 are stored as follows:

BUDU#	(R3) BUD
NE#BUDU#	(R3) BUN

These confixes lead to essentially the same type of analysis as for STAT6, i.e., I is stored in the seventh position of the confix if an infinitive is found, and N, if it is not found. Then the same entries used by STAT6 shift back to the beginning of the word following BUDU# and a match is found on one of the entries:

(R6)BUDBIFB +	WILL#
(R6)BUDBINB +	WILL#BE#
(R6)BUNBIFB +	WILL#NOT#
(R6)BUNBINB +	WILL#NOT#BE

2.1.3. Microglossary Organization

The Russian Master Dictionary was compiled for general-purpose translation rather than for any specific field, in spite of the fact that certain fields of science and technology are better represented in it than others.

In the past, concentrating on a specific field essentially amounted to adding the terms of that field which did not occur in the dictionary. The terms that already existed in the dictionary usually remained unchanged, except in very unusual cases where the existing translation did not render the meaning of the particular Russian word in the given field. In that case, the meaning was changed to include the missing equivalent. As a result of this, and because of other inherent ambiguities in the Russian language, about 2% of the entries in the present RMD are listed with multiple choices as their translations.

Here are just a few typical entries:

NAVARIVANI	COOKING/WELDING	(R3)HR
QEP6	CHAIN/CIRCUIT	(R2)SR
RAK	CRAWFISH/CANCER	(R2)FR
REAKTIVN	REACTIVE/REAGENT	(R7)BU
DIAPAZON	BAND/RANGE	(R1)SR
NAPR4JENI	VOLTAGE/EFFORT	(R3)HR
EST6	IS/EAT	
SUT6	ARE/ESSENCE	
TEXNIK	TECHNOLOGY/TECHNICIAN	(R1)FR
MATEMATIK	MATHEMATICS/MATHEMATICIAN	(R1)FR
SKLEROTIK	SCLEROTIC#PATIENT/SCLERA	(R1)FR
REWENI	DECISION/SOLUTION	(R2)7R

JILY	(DP3)TENDON/VEIN	(R2)OR
SELO	VILLAGE/SAT	

We are not going to discuss the inherent and obvious limitations of the microglossary approach to the solution of the problem of ambiguity in machine translation, but it is quite apparent that if the RMD were segregated into special-field dictionaries, many of the multiple meanings listed above would be eliminated, and the overall readability of output translations would be improved.

So the word QEP6 would have CIRCUITS given as its equivalent in a dictionary devoted to translating electrical engineering and electronics articles, and CHAIN in a dictionary devoted to chemistry or popular journalism.

On a similar basis we could resolve the ambiguity for some of the stems such as: NAVARIVANI, RAK, REAKTIVN, DIAPAZON, NAPR4JENI, and REWENI, although in some cases (especially in the case of NAVARIVANI and REWENI), the elimination of one of the meanings may be disputed.

One must also admit here that for certain fields a word would have to be listed with multiple meanings, even though in other fields only one equivalent would be sufficient, as is the case of RAK, which could be given the translation CANCER in a medical dictionary, but would have to have multiple meanings retained in most of the other dictionaries.

And, of course, there is a whole group of words whose ambiguity could not be resolved by classifying the dictionary into microglossaries. Among this group are EST6, SUT6, JILY, SELO, and even such words as TEXNIKA, MATEMATIKA, and SKLEROTIK.

On the other hand there are many words in the present dictionary, listed with only one equivalent, whose translation could be improved by the microglossary approach, e.g.,

REAKQI	REACTION	(R2)TR
NAKALIVANI	INCANDESCENCE	(R3)HU
REWETK	GRID	(R2)OR
4DR	NUCLE	(R3)BI
IGL	NEEDLE	(R2)OR

So, for instance, in a dictionary tailored for electrical engineering and electronics, REAKQI would be given the translation REACTANCE; for metallurgy NAKALIVANI could be translated an ANNEALING; for mathematics and crystallography REWETKA would have LATTICE as its equivalent; 4DR could be translated as TESTICLE for anatomy or medicine; and IGL could have as its equivalent SPINE for zoology and THORN for botany.

Taking all the pros and cons of the microglossary approach to machine translation, it was decided that this technique of dictionary organization would result in definite improvement in the output translation. Consequently, the work on microglossaries has already been started.

Firstly, a theory of microglossary storage has been developed which will obviate the need of separate discs or bands for each dictionary, and which will allow a complete intermeshing of all the microglossaries in one general dictionary. This technique will be incorporated in the next version of the bidirectional translation system. It may be described briefly as follows:

Each Russian article to be translated will be labeled as to the field to which it belongs (sometimes the title of the journal in which the article appears may be used for this purpose). This label will store a code corresponding to the given field in the computer, which will then be used during the course of translating the article to decide in which microglossary a word that is ambiguous is to be looked up. The microglossaries would consist of words which are ambiguous (such as were discussed above) and which would be prefixed by a confix identifying the given field.

So, for instance we could have:

REWETK	LATTICE	(R2)OR
$^{\rho}\mathrm{MATH}\ ^{\alpha}1$		
REWENI	SOLUTION	(R2)7R
$^{\rho}\mathrm{MATH}\ ^{\alpha}1$		
QEP6	CIRCUIT	(R2)SR
$^{\rho}\mathrm{ELECTR}\ ^{\alpha}1$		
NAPR4JENI	VOLTAGE	(R3)HR
$^{\rho}\mathrm{ELECTR}\ ^{\alpha}1$		
REAKTIVN	REACTIVE	(R7)BU
$^{\rho}\mathrm{ELECTR}\ ^{\alpha}1$		
DIAPAZON	BAND	(R1)SR
$^{\rho}\mathrm{ELECTR}\ ^{\alpha}1$		
RAK	CANCER	(R2)FR
$^{\rho}\mathrm{MED}\ ^{\alpha}1$		
NAVARIVANI	WELDING	(R3)HR
$^{\rho}\mathrm{METAL}\ ^{\alpha}1$		
NAKALIVANI	ANNEALING	(R3)HU
$^{\rho}\mathrm{METAL}\ ^{\alpha}1$		

REWETK $\rho_{CRYST}\ \alpha_1$	LATTICE	(R2)OR
4DR $\rho_{ANAT}\ \alpha_1$	TESTICLE	(R3)BR
RAK $\rho_{BOT}\ \alpha_1$	CANCER	(R2)FR
IGL $\rho_{BOT}\ \alpha_1$	THORN	(R2)OR
IGL $\rho_{ZOOL}\ \alpha_1$	SPINE	(R2)OR
RAK $\rho_{ZOOL}\ \alpha_1$	CRAYFISH	(R2)FU

and so forth.

All of these entries would also have to be stored under a confix, say $\rho\nu$, which would match on any of the given field-codes, as follows:

REWETK $\rho_{\nu}\ \alpha_1$	GRID	(R2)OR
REWENI $\rho_{\nu}\ \alpha_1$	DECISION/SOLUTION	(R2)7R
QEP6 $\rho_{\nu}\ \alpha_1$	CHAIN	(R2)SR
NAPR4JENI $\rho_{\nu}\ \alpha_1$	EFFORT	(R3)HR
REAKTIVN $\rho_{\nu}\ \alpha_1$	REACTIVE/REAGENT	(R7)BU
DIAPAZON $\rho_{\nu}\ \alpha_1$	RANGE	(R1)SR

	RAK	CRAYFISH/CANCER	(R2)FR
ρ_ν α_1			
	NAVARIVANI	COOKING	(R3)HR
ρ_ν α_1			
	NAKALIVANI	INCANDESCENCE	(R3)HU
ρ_ν α_1			
	4DR	NUCLE	(R3)BI
ρ_ν α_1			
	IGL	NEEDLE	(R2)OR
ρ_ν α_1			

As can be seen, this technique is very flexible, and would lend itself very well to future changes in field classification. It also has the important advantage of utilizing the general dictionary, with its wealth of lexical data, for translating any desired field. For unlisted articles, the meaning of ambiguous words would be gotten from the $\rho\nu$ table.

Plans were made and the work has already started on reorganizing the RMD along the lines discussed above, and for this purpose a list of 13 very basic and useful (from the standpoint of AF requirements) fields has been established.

Following is a list of these fields together with the codes that will be used for their identification:

Aeronautics and Aviation	A
Biology	B
Chemistry	C
Mechanical Engineering	D
Electrical Engineering and Electronics	E

Metallurgy	F
Military Sciences	G
Medicine	H
Mathematics	M
Nuclear Physics	N
Physics	P
Space	S
Earth Sciences	T

It would be a simple matter to further expand this list, as all that would have to be done to the dictionary is to enter the affected words under the new confix representing the added field.

To aid us in the future studies of microglossary construction and classification, a complex concordance program was written for the 7090 computer. The program has the following options:

1. It will produce a complete concordance of a given text This means that the sentences of the text will be numbered, and all the occurrences of any given word will be collected and printed together. Under each heading, or "study word," the total frequency of occurrence of that word in the particular text is given. The actual study word is signaled in each of the sentences in which it occurs by printing two asterisks just before the study word itself. A sentence is printed only once with any particular study word, although the study word may occur more than once in some sentences; each occurrence of the study word is signaled, however, even when it occurs more than once in a sentence.

2. The program will yield, as an option, a "directory" concordance of a text. Here, instead of printing each sentence in which each study word occurs, there are two separate but considerably smaller outputs. The first is a single copy of the text, sentence by sentence, with the sentences numbered as they are for concordance purposes. The second output is the "dictionary," an alphabetic list of the study words, with the frequency of occurrence given as above; but instead of the complete sentences in which each word appears, only the numbers of the sentences are given. Since 19 sentence-numbers can be printed on a line of paper, while printing the actual sentences may take several lines for each, it is clear that the volume of output from this option is manyfold less than from option 1. For a million-word text the directory alone may be a foot-high stack of wide printing paper; the full, option 1, concordance for such a text would probably be around 20 times as bulky. The saving in machine-time is equally great.

3. The user may then request a complete concordance of specific study words in the text. It appears more useful to produce, for example, a special concordance of a specific list of prepositions, bound, labeled, and kept separate, and then another concordance; say, of reflexive verb-forms, than to request the complete concordance of every word in a very large text. This option is available with either option 1 or option 2; that is, one does not need a directory of a text in order to request a full concordance of any particular items.

4. This option is the complement of option 3: in option 4 one requests a concordance (either option 1 or option 2 type) minus specific words. It is assumed, for example that in a concordance of English-language text the user will not want a listing of all the sentences containing "the," "a," or "and"; in option 4 these, and other words not wanted, are specified, and are then left out of the concordance.

5. On top of any of the above options one can request that "endings" be removed from actual text words, and the remaining "items" concorded together. Thus; in English, one may ask that final -s be removed from all words, thus presumably bringing together singular and plural nouns, or plural and singular present-tense verbs. One runs the risk of putting together items which do not linguistically belong together, but the technique turns out to be very useful in practice.

All the above options are programmed and operating. We have approximately 7,000,000 running words of Russian technical text punched on paper tape. This text has been put on magnetic tape, and work has begun on concording all of it, separately for each individual technical field. When the directory concordances have been made for all the text it is planned to merge them, giving a single large directory for 7,000,000 words. This will be used for many purposes, among them microglossary studies, studies toward a semantic classification of the dictionary, and checking the completeness of the RMD.

In addition to this, a program has been written to extract from the directories those study words above a given -- optional -- frequency of occurrence. These extracted study words are then sorted on the endings, a technique which in Russian brings together syntactically similar items: because of the high degree of inflection in Russian, words with the same characters at the end are very likely to be of the same part of speech. A relatively small amount of human intervention is then necessary to separate out the members of any such group which do not linguistically belong there. Another program then removes the endings from the full study words, producing stems. This ability is useful in the production of special dictionaries, or in updating an existing dictionary on the basis of lacunae actually found by testing with real text.

2.1.4. Multipass Routines

The control section which contains the analytic routines of the Multipass dictionary has been augmented to 25,000 entries. For the sixteen-pass translation system, passes zero, 1, 2, 3, 4, 5, 6, 8, 9, 10, 12, 15 and 16 have been completed and demonstrated with a limited vocabulary which nevertheless thoroughly tests the feasibility of the system. The vocabulary of approximately 800 Russian stems is represented by 900 Search Input State (SIS) entries and 2000 Pass 16 English readout entries. The increased number of SIS entries is due to "longest match" considerations and the absence of Process State routines to perform morphological derivation of parts of speech. The increased number of English readout entries provide the improved output translation of Russian stems by providing alternate translations according to the "meaning preference" tags in the Process Word format; this tag is updated on the basis of syntactic and semantic linkages established between the constituents of the sentence, and the updated tag represents a reduction in the semantic ambiguity of an input Russian stem.

The 25,000 entries in the control section constitute the routines which perform the syntactic and semantic analysis of input sentences by table lookup. The sentence structure is determined to the degree necessary to reduce grammatic and semantic ambiguity, to perform word re-arrangement, and to make insertions of English connecting words (such as "of," "by," "to") or English verbal auxiliaries to form compound tenses.

The Multipass routines consist of elaborate tables which match and link noncontiguous elements of the sentence. Both syntactic and semantic linkages are established by the appropriate routines. Some basic programming techniques have been developed and are applied universally in routines of the various passes. Representative routines and programming techniques which are now in practice are enumerated below.

Transliteration - When the SIS entries fail to find an input word (either the stem or the ending), the entire word is transliterated in the output. Numbers or other symbols interspersed in the text are treated similarly. Transliteration requires two steps. An SIS routine determines the length of the input word and transcribes it character for character into a variable-length Process Word with special part-of-speech tags. Ordinarily, only limited possibilities exist for the analysis and linkage of such Process Words during the subsequent Process State processing. During the English readout stage of Pass 16, another routine recognizes the part-of-speech tags and proceeds to transliterate the input characters into their output equivalents.

Establishment of Linkages - A basic programming technique is common to the establishment of all linkages. The routines which establish the different types of linkage differ only in the parameters which are matched by the entries. For economy in the matching entries, two stages of lookup are customarily used. The first stage performs a preliminary scan for some basic pattern (of parts of speech, for example) which may involve either two or three elements of the sentence. The preliminary scan proceeds either to the left or to the right from any chosen word and locates the Process Words whose parameters (part-of-speech tags) satisfy the necessary conditions for the linkage. Intervening words are selectively skipped as desired. Having found the pattern, this stage sets index registers to indicate the locations of the linked Process Words. Typical entries are shown below. The parameters of the example are part-of-speech tags.

(1) Locate an "NN" to the left:

```
(P2)B + *ØØNN          τ  (EC)  μ  C
(P2)B + *              τ  (EA)
(P2)B + *ØØØ (PHI)     τ        μ  AB
```

If the Argument Index Register (AIR) is now pointing to an "NN" the first entry matches, and the "(EC)" instructions set the Continuation Index Register (CIR) at the location of the Process Word on the right, in preparation for the next search among entries which are prefixed with "(P2)C" in paragraph (2). If the current Process Word is not an "NN", the (EA) instruction in the second entry sets

AIR to the location of the Process Word on the left, and the above procedure is repeated. If the beginning of sentence is encountered before an "NN" is found, the third entry matches a special Process Word containing a "(PHI)" symbol and causes an exit to the "(P2) AB" routine (not part of this example).

(2) (a) Locate an "NA" to the right of the "NN":

(P2)C + * ØØNN(DC)*ØØNA	T		μ D
(P2)C + *	T	(EA)	μ B

If the "NN" is followed immediately by an "NA," the first entry matches and causes an exit to the "(P2)D" routine for further analysis; the (DC) instruction in the argument utilizes the pre-set address in the CIR register and causes immediate advance to that location. Thereby, the parameters in two contiguous Process Words are matched in one lookup. If this match fails, the second entry returns to "(P2)B" to locate another "NN" to the left, as in paragraph (1).

(2) (b) Locate an "NA" to the right of the "NN" with skipping permitted over only "X" and "Q":

(P2)C + *ØØNN(DC)*ØØNA	T		μ D
(P2)C + *ØØNN(DC)*ØØX	T	(DC), (EC)	
(P2)C + *ØØNN(DC)*ØØQ	T	(DC), (EC)	
(P2)C + *	T	(EA)	μ B

The first entry is identical to that of the preceding example; however, because of the effect of the second and third entries, the (DC) instruction now causes the linkage of "NN" and "NA" items in noncontiguous, as well as contiguous Process Words. If an "NA" is not immediately to the right but either "X" or "Q" is, the second or the third entry matches. In the function field, the (DC) instruction causes an advance to the location of the "X" or "Q" (just as the (DC) instruction in the argument did). Then the, "(EC)" instructions set the CIR register to the location of the Process Word to the right of that point, with the result that noncontiguous words will be processed if the contents of CIR are used. Because no change in confix is specified, the next lookup will again involve the above entries prefixed with "(P2)C". If "X" or "Q" Process Words occur in succession, they will be skipped successively by recursive matches until either the desired linkage is made by a match with the first entry or else skipping is terminated because of a part-of-speech tag other than "X" or "Q." The fourth entry terminates skipping and returns to "(P2)B" in order to locate another "NN" and restart the entire procedure from that point.

When the elements of the desired linkage have been located by a procedure such as the example above, the second stage of the routine performs detailed analysis of the selected Process Words. Because the contents of AIR and CIR (and another index register, MSKR) are available, this analysis is effective regardless of the positions of the words in the sentence. The analysis is performed by means of tabular entries which enumerate all desired combinations of grammatical, semantic, or processing tags in any of the selected words. One entry in the table matches the given configuration and thereby identifies the linked items in detail. The result of the analysis is recorded by updating the tags in any or all of the linked words according to the degree of ambiguity which has been resolved. This information is then available in the Process Words for still further analysis during subsequent passes. The information is highly codified according to a standard format and convention. Typical entries for such analysis are given below. This stage of the routine is entered by changing the secondary confix symbol to "D" as was done by those previous entries containing "μD."

```
(P2)D + *00NN(DP17)DXC(DC)*00NA(DP18)TVT(DP7)X(DC)(DP7)BμE
(P2)D + *00NN(DP17)DXD(DC)*00NA(DP18)OST(DP7)X(DC)(DP7)DμE
(P2)D + *00NN(DP17)IXJ  (DC)*00NA(DP18)PTT(DP7)A(DC)(DP7)FμE
(P2)D + *00NN(DP17)DXF(DC)*00NA(DP18)OST(DP7)X(DC)(DP7)DμE
(P2)D + *00NN(DP17)IXJ  (DC)*00NA(DP18)QFT(DP7)A(DC)(DP7)CμE
(P2)D + *00NN(DP17)IX1  (DC)*00NA(DP18)OST(DP7)X(DC)(DP7)DμE
(P2)D + *00NN(DP17)DXL(DC)*00NA(DP18)OST(DP7)X(DC)(DP7)DμE
(P2)D + *00NN(DP17)DX1 (DC)*00NA(DP18)DQT(DP7)X(DC)(DP7)XμE
(P2)D + *00NN(DP17)IX6  (DC)*00NA(DP18)IPT (DP7)X(DC)(DP7)XμE
(P2)D + *00NN(DP17)DX1 (DC)*00NA(DP18)TVT(DP7)X(DC)(DP7)FμE
(P2)D + *00NN(DP17)DX6 (DC)*00NA(DP18)OST(DP7)X(DC)(DP7)DμE
(P2)D + *00NN(DP17)IYG (DC)*00NA(DP18)VWT(DP7)A(DC)(DP7)DμE
(P2)D + *00NN(DP17)IYG (DC)*00NA(DP18)MPT(DP7)D(DC)(DP7)AμE
```

Word Order Routine - This routine in Pass 16 establishes the order in which Process Words are to be translated into English on the basis of word re-arrangement tags $r_1 r_2$ which appear in the Process Word format. These tags are filled in during the preceding Process State analysis. The sequence of re-arrangement tags, from each Process Word to the next, indicates linked word groups or phrase structure within the sentence. The present routine accomplishes only first-order re-arrangement; that is, the arrangement of relatively major phrases as specified by the r_1 tag. Second-order re-arrangement, specified by $r_1 r_2$ together, would provide also for the re-arrangement of words or subphrases within a major phrase. For only first-order re-arrangement, the r_2 tag is always "." and the r_1 tag has one of the following codes:

"."	This word remains in its relative position.
"P"	A word or group to the right with tag "N" is to precede this word (or group) with tag "P".
"N"	This word and all successive "N" words are to be translated before the "P" words on the left.

Thus, as a result of prior analysis, the following configuration of r_1 tags might appear in a sentence, and the corresponding Process Words would be translated in the order indicated:

(bos)	.	.	P	P	N	N	N	.	P	N	N	P	P	P	N	(eos)
	1	2	6	7	3	4	5	8	11	9	10	13	14	15	12	

Closely associated with the word-order routine is the word-insertion routine which examines the Insertion tag in each Process Word before that word is translated. The value of this tag determines the insertion to be made in the English output.

2.1.5. Conversion of RMD to Multipass Entries

The Russian Master Dictionary has undergone considerable revision and updating to improve the lexicon for Multipass usage as well as for Bidirectional Single-Pass usage. Further classification of RMD entries has been accomplished by the application of new computer processing techniques. These efforts are directed toward organization of the lexical data in a form suitable for the automatic generation of full-scale Russian Multipass dictionaries in a truly general way. Heretofore, only the "backup" entries for Pass 16 English readout could be produced from RMD by computer without any human intervention.

Each RMD entry consists of four fields: Acquisition Number, Argument (Russian), Function (English), and Confix (grammatical classification). This format is primarily oriented toward Single-Pass translation (including Bidirectional). As an implement to converting the RMD entries to Multipass entries, an intermediate file "RMDV" has been created by annexing to each entry a fifth field, "Field V," for general Multipass use. The overall purpose of Field V is to collect explicit codified information which will control a computer program to generate the appropriate Multipass vocabulary entries in a straightforward pass over the RMDV file.

In establishing Field V, an improved system for the assignment of semantic tags to Russian stems has been implemented by a computer program. The semantic tags will appear ultimately within the Process Word format as a unique, coded representation of each Russian stem. The improved semantic classification of stems is based on codifying the root and its affixes (if any) independently. By this procedure, those RMDV entries whose arguments are morphological variations of the same root are correlated by the assignment of similar semantic tags.

The computer program processes each Russian stem in the RMD file, and by applying basic morphological rules for the derivation of word forms, it presents a hypothetical trisection of the stem into its root, prefix, and suffixes. This information is recorded in Field V. With a minor amount of automatic recoding, this information will constitute the semantic portion of the multipass entries to be generated. In the course of processing the Russian stems, the program also records, in Field V, a codified trace of its activity in order to classify the stems according to the applicable morphological rules. The latter information, when recoded automatically and supplemented manually, will serve to control the subsequent RMDV-to-Multipass conversion program. The program which created Field V includes the following features:

a. Common linguistic prefixes are isolated from the root. Normally, the longest possible prefix is removed; for example, prefix PREDO has preference over PRE, and OB over O. In Field V, equivalent prefixes are represented by the same code; for example, OB and O are both codified as OB. By referring also to approximately 1500 common Russian roots, the program performs this task quite accurately. For example, the root VOD is not dissected as VO-D, and the word OBEGAT6 is dissected properly as O-BEGAT6 rather than as OB-EGAT6.

b. Common derivational suffixes are isolated from the root according to grammatical classifications. The suffixes are identified, and a codified trace of the formation of the stem from the root is recorded in Field V. This analysis is not restricted to the 1500 common roots.

c. Endings are truncated from full-word entries whose purpose in the dictionary is to prevent false shorter matches. These entries are identified by the occurrence of a (DPn) instruction in the Function field.

d. The program correlates pairs of entries which differ by a "mobile vowel."

e. The program correlates the root forms which differ morphologically by a change in the final consonant. For example, the root ALC is transformed into ALK when recorded in Field V. By a similar transformation, the tort/tolt groups are identified and correlated; for example, GOLOV is transformed into GLAV by the program. Among the 1500 common roots, even more extensive correlation is performed by assigning identical root codes in Field V. Thus, the alternate root forms DR, DER, DOR, DAR, DIR, DYR are correlated.

f. When a stem is derived from either of two homographic roots, the program examines the morphemes adjacent to the root to resolve the ambiguity. After this analysis the roots, homographic in Russian, are distinguished in Field V by the assignment of distinct semantic codes. For example, Field V distinguishes, with fair reliability, between stems derived from VOD "water" and those derived from VOD "to lead." This analysis is applicable only to the common roots furnished to the program.

EXCERPT FROM THE RMDV FILE SHOWING CORRELATION BY SEMANTIC TAGS

MNOGOVODN	abounding in water	(R7)BU	)VODN)XM
MNOGOVODNOST	abundance of water	(R2)VU	)VODN)XM-OST
BEZVODEN	anhydrous	(R7)BU	)VODN)BZ
BEZVODN	anhydrous	(R7)BQ	)VODN)BZ
OBVODN	surrounding	(R7)HO	)VODN)OB
OBVODN4	irrigat	(R4)BE	)VODN)OB
OBVODN4T(PN)	(DP6) irrigat	(R5)CE	)VODN)OB
OBVODNI	(DP6) irrigat	(R5)CE	)VODN)OB
OBVODNH	(DP6) irrigat	(R5)CE	)VODN)OB
OVODN	irrigat	(R4)BE	)VODN)OB
OBVODNEN	irrigated	(R7)VU	)VODN)OB-N
OBVODNENI	irrigation	(R3)HU	)VODN)OB-ENI
OVODNENI	irrigation	(R3)HU	)VODN)OB-ENI
OBVODNITEL6N	irrigating	(R7)EU	)VODN)OB-TLN
PODVODN	underwater	(R7)HU	)VODN)PD
PODVODNI	submariner	(R3)DR	)VODN)PD-NIK
PEREVODN	translated/transferred	(R7)HU	)VODN)PE
PEREVODNIK	adapter	(R1)SR	)VODN)PE-NIK
VVODN	introductory	(R7)EU	)VODN)WO
VYVODN	discharge	(R7)HU	)VODN)WY
VZVODN	platoon/cocking	(R7)HU	)VODN)WZ
ZAVODN	clockwork	(R7)HU	)VODN)ZA
ZAVODN4	flood	(R4)BC	)VODN)ZA
ZAVODNI	(DP6) flood	(R5)CC	)VODN)ZA
ZAVODNH	(DP6) flood	(R5)CC	)VODN)ZA
ZAVODNEN	flooded	(R7)VU	)VODN)ZA-N
ZAVODNENI	flooding	(R3)HU	)VODN)ZA-ENI
NADVODN	above water	(R7)EU	)VODN)ND
NAVODN4	(DP6) flood	(R5)BC	)VODN)NA
NAVODNI	(DP6) flood	(R5)CC	)VODN)NA
NAVODNH	(DP6) flood	(R5)BC	)VODN)NA
NAVODNEN	overflowed	(R7)VU	)VODN)NA-N
NAVODNENNOST	overflowing	(R2)VU	)VODN)NA-NNOST
NAVODNENI	flood	(R2)7R	)VODN)NA-ENI
PROVODN	wire	(R7)HU	)VODN)RO
PROVODNI	conductor	(R3)DR	)VODN)RO-NIK
RAZVODN	separating	(R7)HU	)VODN)RZ
SVODN	summary	(R7)EU	)VODN)SO
SVODNICA	procur	(R4)AE	)VODN)SO-NICA
SVODNICESTV	procuration	(R3)FU	)VODN)SO-NICESTV
SVODNICESK	pandering	(R7)EU	)VODN)SO-NICESK
(R3)VBV+VYVODIL	(DP6) led out/concluded	(R6)VBV	)VOJ)WY
VYVODIMOST	derivabilit	(R2)SY	)VOJ)WY-MOST
NIZVODIVW	bringing down	(R7)WU	)VOJ)NI-VW
PROVODIMOST	conductance	(R2)VU	)VOJ)RO-MOST
PRIVODIMOST	reducibility	(R2)VU	)VOJ)RI-MOST
SVODIMOST	reducibilit	(R2)SY	)VOJ)SO-MOST
ZAVODIWK	factor	(R2)BY	)VOJ)ZA-WK
ZAVODIWEK	factories	(R2)BU	)VOJ)ZA-WK
OTVODITEL	diverter	(R1)SR	)VOJ)OT-TL
PREDVODITEL	leader	(R1)AR	)VOJ)RD-TL

This excerpt from the RMDV file shows how the semantically related stems are associated by the computer program. Not shown are the Acquisition Numbers of the entries or the highly codified derivational tracing information. With reasonably good reliability, semantically equivalent stems have been codified identically and uniquely, while morphologically related entries have received similar semantic codes. Although this task is not entirely computable, further mechanical processing of the RMDV file, with human intervention, will quite efficiently refine these perfunctory results. The juxtaposition of related entries in the file enhances the feasibility of mechanically discovering inter-entry relationships. Such relationships will assist in the refinement of RMDV and will also be exploited in generating a highly efficient multipass dictionary.

After refinement of RMDV, the suffix information will be compressed to two-character codes, and the tracing information (not shown above) will be converted into control information for the RMDV-to-multipass conversion program. The eventual format of Field V will be as follows:

rrrr pp ss cccccc

where rrrr is the root code

pp is the prefix code

ss is the suffix code

cccccc is the conversion control information.

The first eight characters constitute the semantic tags to be substituted directly into multipass vocabulary entries.

A typical multipass vocabulary entry to be generated for the Search Input State is as follows:

RUSSIAN τ(F--), (DPn) rrrrppss . .*(B--) μ $\rho\rho$

During Mark-II operation, this entry matches a Russian stem and transcribes it into an intermediate-language Process Word in the Process Region. Initially this Process Word consists only of the forward and backward length tags and the semantic encoding of the Russian stem. The confix $\rho\rho$ directs the subsequent search to look up the remainder of the Russian word in a common set of endings tables; this lookup operation supplies part-of-speech tags and other codified grammatical information.

Further updating of the Process Words is performed during the Process State analysis of the entire sentence. Elaborate tables of control entries are searched to link the constituents of the sentence and to determine sentence structure to the extent necessary to resolve semantic and grammatical ambiguities of the individual Process Words.

As the final step, the English translation of each Process Word is obtained by searching the vocabulary entries of the Process State. A typical such entry is as follows:

(P 16) Ø 1+*ØØqq(DPn) rrrrppss x T, (EA) (OD) English μ,

where qq represents part-of-speech tags, x is a possible meaning preference tag, and (EA) and (OD) are Mark-II instructions. The confixing provides for appending English suffixes, word insertion, and word re-arrangement by the appropriate Process State routines.

Although the prior state of the RMD file would have permitted the automatic generation of similar entries with arbitrary semantic tags, this task was postponed in view of the significant improvement in organization achieved by the implementation of Field V. With Field V, inter-entry relationships can be discovered and exploited by mechanical means. This will afford more economical processing of the lexical data and will produce more compact and efficient multipass dictionaries.

Prior methods of generation would have converted every RMD entry into at least two multipass entries (one Search Input State and one Process State "backup" entry). Additional Process State vocabulary entries to encompass meaning preference would necessarily have been prepared manually. However, the cccccc data in Field V will serve to suppress the generation of extraneous entries or to generate additional entries as required (such as for meaning preference). Conditional choice in the format of generated entries is also provided. For example, whenever several entries have identical semantic tags (after refinement of RMDV), each RMDV entry may require conversion to a Search Input State entry, but a single Process State entry will suffice for English readout. Field V will also indicate the degree to which derivational analysis can be performed on the Mark II; in this case, certain SIS entries can be suppressed when the appropriate derivational routines are included in the control dictionary.

The cccccc data, by its nature and method of derivation, will tend to make explicit the purpose of each argument in the RMDV lexicon. By changing interpretive parameters in the RMDV-to-Multipass conversion program, complete flexibility is possible. Thereby, the conversion program is not restricted to any particular multipass translation process nor to any specific multipass dictionary organization. New techniques of analysis or translation by table lookup can be readily embodied in a full-scale multipass dictionary.

2.2 Linguistic Research

2.2.1 Housekeeping

2.2.1.1 Introduction. Practical machine recognition of the syntactic structure of random, natural language sentences is not a straightforward task that can be accomplished by the use of rudimentary linguistic and programming techniques. The most sophisticated and powerful linguistic theories must be applied; and, since at present the pertinent area of linguistic theory is undergoing extremely rapid development, application unavoidably involves evaluation, refinement, and extension of the theory. Moreover, if the goal is a practical and economical recognition system rather than a demonstration "in principle," the programming must be specially tailored to fit the linguistic formalism.

If the problem were less complex than it actually is, or if some of the complexities could be permanently ignored, there would be no need to devote special attention to communication between the linguist and the programmer. Each could be more or less independent of the other. The linguist could state his rules without worrying about whether they were programmable (in a practical way) and the programmer could program the linguist's rules without very much understanding of their underlying structure. Of course, such understanding by the programmer is always desirable, but need not always be essential. However, the crucial need for the development of effective communication between the linguist and the programmer in language data processing is amply demonstrated by the fact that, at present, despite several years of serious effort to incorporate the relevant achievement of linguistic science into MT programs, there remains large areas of sound accomplishments in grammatical description (particularly at the sentence level) which seem to be too complex for inclusion in current programs.

The problem of linguist - programmer communication has of necessity been faced by every MT group in the country and has received a fair amount of discussion in the literature, at conferences, etc. However, the problem has not been formalized. Thus, each individual group can simply describe its procedure, with perhaps some discussion of evolution, motivation, special features, etc., but cannot characterize it in terms of some generally accepted classification framework or set of categories.

The existing procedures fall roughly into three groups. In one group, there is little explicit attention to linguist - programmer communication. There will of course be some grammatical coding conventions and there may be some flow-charting or rule-writing conventions, but for the most part, the linguist prepares his material in whatever form suits him, and then attempts, through informal discussion, to explain to the programmer as much as he seems to need to know. This procedure may seem to work at first, but the usual result is that after awhile the linguist is surprised to find that minor linguistic modifications or additions entail a major programming effort or, even worse, cannot be done practically within the existing program. Finally, the program becomes an unwieldly patchwork, becomes unmanageable, and must be completely rewritten from scratch. The objection to this procedure is not that programs must be radically revised or completely rewritten periodically, but rather, that the evolution of the programming is not controlled by the evolution of the linguistic conceptions.

In the second group, linguist-programmer communication is effected by making the programmer dominant; i.e., practical (or at least feasible) programming is ensured by selecting particular programming devices (for example, pushdown storage). The linguist is then forced to limit himself to statements of a form permitted by the particular programming devices selected. Again, the objection to this procedure is not that particular programming devices are selected (in fact, one of the main purposes of explicit attention to "housekeeping" is to provide a sound linguistic basis for such selections), but rather, that the criteria for selection are dominated by programming convenience rather than linguistic necessity.

In the third group, an attempt is made to eliminate the programmer entirely by setting up a special linguist's programming language. Initially, programmers are needed to write interpreters and compilers for the special programming language, but once the job is done, the programmer is no longer needed, and presumably the linguist effectively has direct access to the machine. This approach has some merit and will probably be useful in the future (not the immediate future), but it is completely useless for present production purposes and is at best of only marginal suitability for current research purposes. The idea of an explicit special linguist's programming language is of essential significance (and is also one of the main purposes of explicit attention to housekeeping) but the attempt to eliminate the programmer is premature. A special linguist's programming language must be based on an explicit statement of the types of structure which the linguist uses in linguistic description, and on the types of manipulations which are performed on these structures either as an essential part of linguistic description or as a part of mechanical sentence analysis. Such a programming language will be useful to the linguist just to the extent that it is restricted to express no more and no less than the linguist needs. But this area of linguistic science is undergoing extremely rapid (in fact, revolutionary) development at present, which means that the more useful (through suitable restrictions) a programming language is to the linguist, the more rapidly it will tend to become obsolete. If the programming language is continually modified, then programmers will continually be needed to make the necessary changes in the interpreter and compiler associated with the programming language. Thus, in order to eliminate the programmer for some reasonable length of time, the programming language would have to be made more general (less structured) than necessary. But then, in order to express linguistic structure, the linguist must in effect add structure to the language; i. e., he must do some programming. Thus, the programmer has not been eliminated, but rather, the linguist has partially been turned into a programmer. This not only hampers the linguist, but also usually results in programs which require much more machine running time than necessary.

The above discussion of the strengths and weaknesses of current approaches to linguist-programmer communication is intended to emphasize the importance of the problem and to show that further research in this area will certainly provide more economical techniques in language data processing, and may in fact be essential if the machine techniques are to be kept up-to-date with respect to developments in linguistics or are to contribute to such development.

Before proceeding to the discussion of the approach proposed herein, it will be useful to attempt a general statement of the problem.

From the point of view of the linguist (or almost any user, for that matter) the general-purpose computer is essentially unstructured (hence the name "general-purpose" computer). None of the individual machine instructions (CLA, TIX, etc.) have any linguistic significance. Each elementary linguistic operation such as "rewrite X as Y + Z keeping track of the fact that this particular Y + Z is an expansion of this particular X, " or "scan the sentence to the left for an occurrence of word class U but don't pass over any occurrences in word classes V or W, or "test item A and B for grammatical agreement, " etc., requires a large number of appropriately selected and ordered machine instructions for its execution, i. e., requires a more or less complicated program written in machine language. A set of elementary linguistic operations could have a corresponding set of machine language programs.

However, it usually turns out in practice that many of the machine language programs in the set would have one or another part in common. Therefore, in order to achieve optimum programming economy, the set of elementary linguistic operations is realized not by a corresponding set of machine language programs but by a program consisting of a set of subprograms none of which corresponds to any of the elementary linguistic operations. This program in effect adds structure to the computer. However, despite efforts to achieve optimum programming economy, the structure added to the computer by the program may not correspond very closely to the syntactic structure with which the linguist operates. The fact that a program "works" proves nothing. For example, if airline routes were arranged in such a way that to travel between any two cities in the United States, one had to go by way of the North Pole, the system would "work" in the sense that travelers would reach their destination, but it would be obvious that the airline route structure did not correspond very closely to the structure of optimum passenger traffic. In the case of language data processing, the lack of close correspondence is usually not obvious until the linguist tries to modify his system and finds very little correspondence between the magnitude of a linguistic modification and the magnitude of the corresponding modification of the program or the magnitude of the effect on machine running time.

Summarizing, our objective in general terms is to develop a procedure for linguist-programmer communications which will enable us to achieve the desired correspondence between linguistic and program structure.

2.2.1.2 Preliminary Investigation. As a first step, the linguist must make an explicit statement regarding:

1. The type of structures to be represented,
2. The manipulations to be performed on these structures,
3. The search strategy.

This statement should not be a description of a particular recognition scheme adapted to a particular language, but rather, should be sufficiently general to be adaptable to any language. It is important to note that the aim is not some absolute notion of maximum generality (a Turing machine or a general-purpose computer or the axioms of set theory) but a restriction to just the generality needed for linguistic description. If the statement is too general, part of the programming will have to be done, in effect, by the linguist. If the statement is too restricted, the linguist will have lost some necessary descriptive power.

Some preliminary work has already been done on parts (1) and (3) of the linguistic statement. The work on part (3) is described in the section of this report entitled Search Strategy. A description of the work on part (1) is given below.

The basic type of structures considered thus far can best be described as ambiguous trees. In the definition of a tree, a node has only one node above (i.e., only one branch entering it) and any number of nodes below (branches leaving it). An unambiguous sentence or one particular interpretation of a syntactically ambiguous sentence can be represented by such a tree. See, for example, Figure 2.2-1 which illustrates the representation of a Russian sentence in terms of an oversimplified grammar containing only four kinds of syntactic relations (P-predication, C-complementation, M-modification, J-conjunction). The dotted lines in the figure connect pro-elements and their antecedents but are not to be considered part of the tree (in fact, if they were, the representation would no longer be a tree). This particular example of sentence structure representation has some special features (such as the way the J-nodes are used and the representation of functives) which are not relevant to the present discussion. A more conventional and extensive grammar of Russian is shown in Figure 2.2-2.

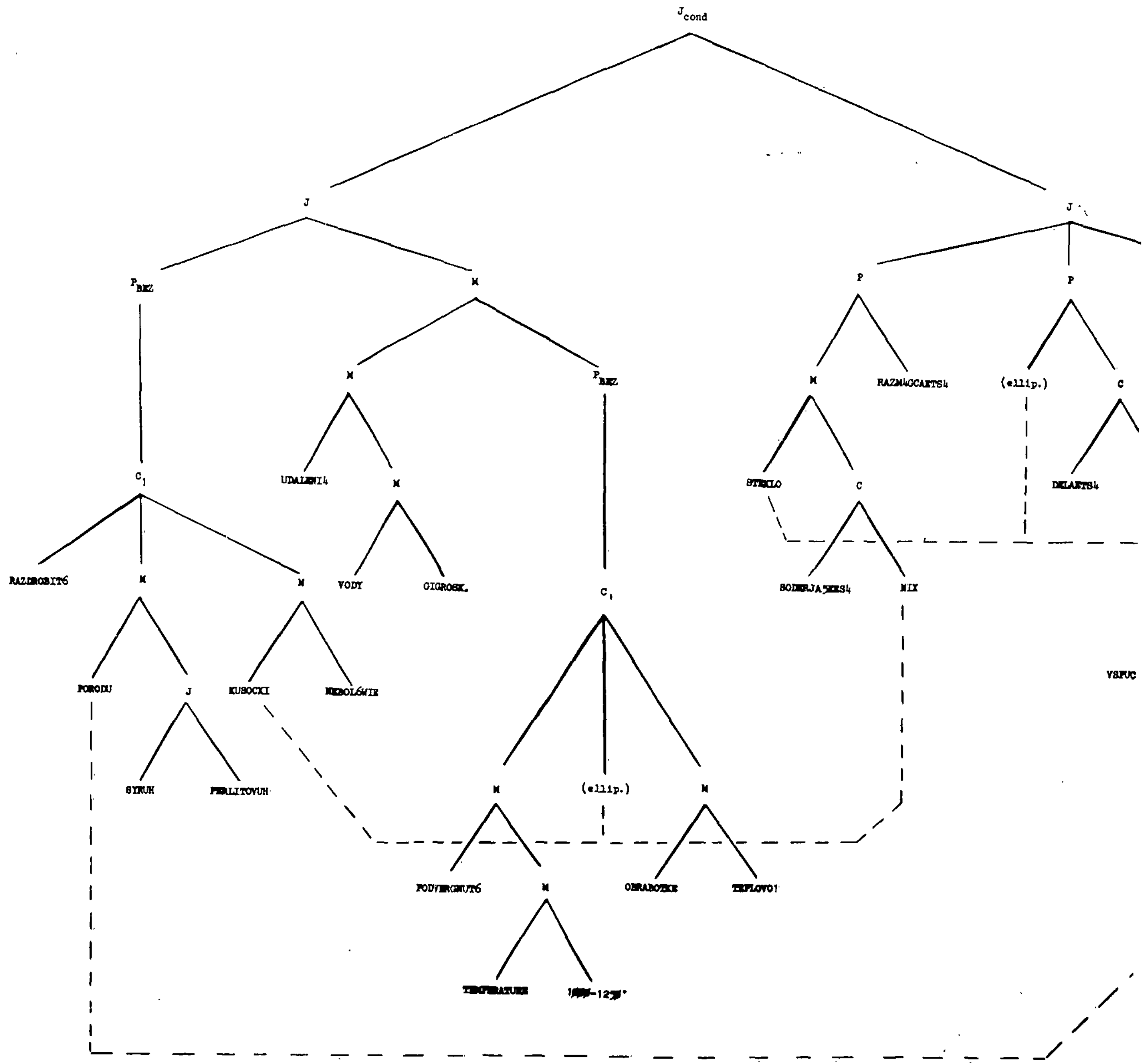
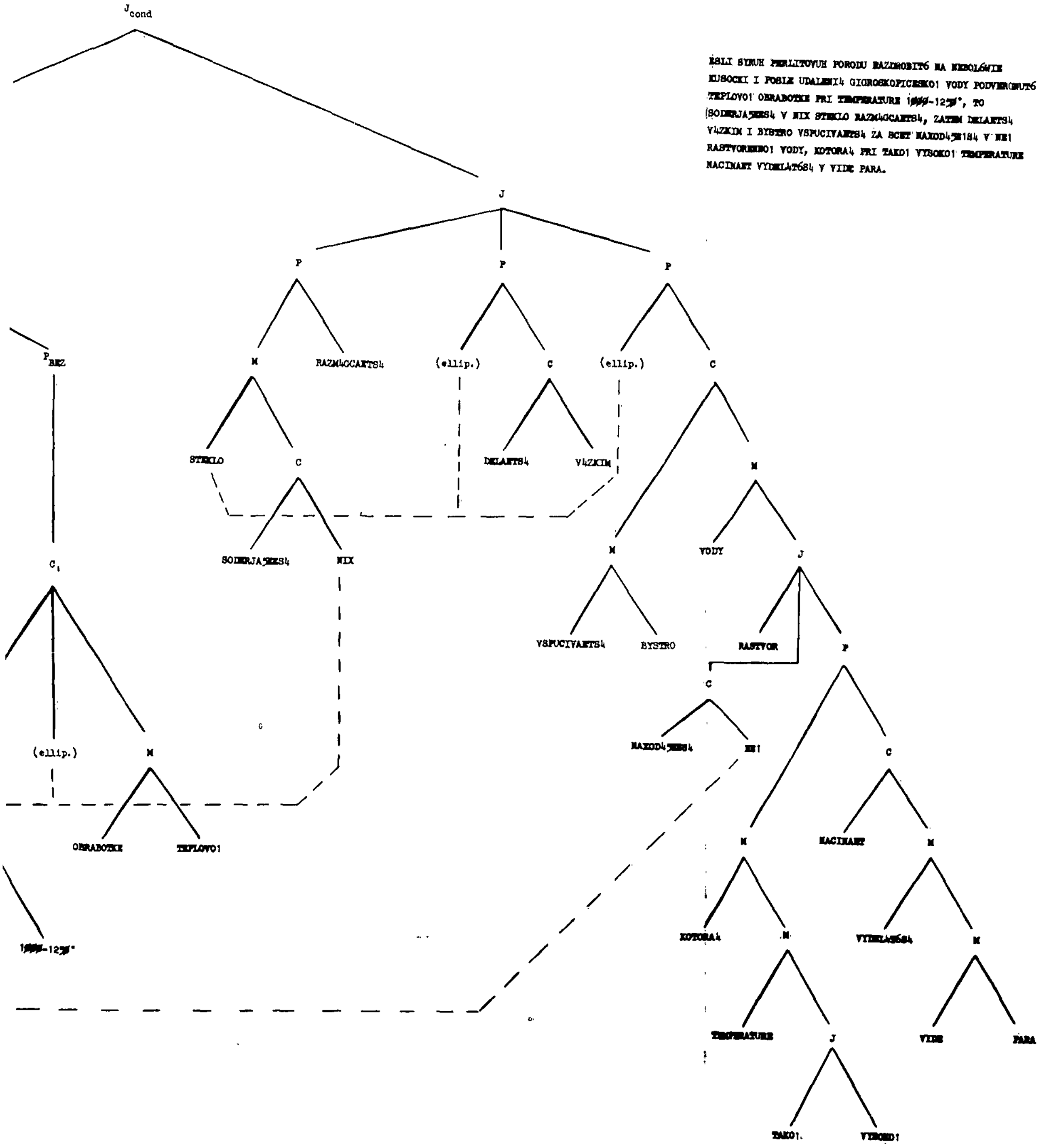


Figure 2.2-1 Diagram of a Russian Sentence in Terms of a Sim

ESLI SYRUH PERLITOVUH PORODU RAZDROBIT6 NA NEBOL6WIE KUSOCKI I POSLE UDALENI4 GIGROSKOPICESKO1 VODY PODVERGNUT6 TEPLOVO1 OBRABOTKE PRI TEMPERATURE 1000-1250°, TO SODERJA5ES4 V NIX STEKLO RAZM4GCAETS4, ZATEM DELAETS4 V4ZKIM I BYSTRO VSPUCIVAETS4 ZA SCET NAXOD4518S4 V NE1 RASTVORENNO1 VODY, KOTORA4 PRI TAKO1 VYSOKO1 TEMPERATURE NACINAET VYDEL4T6S4 V VIDE PARA.



agram of a Russian Sentence in Terms of a Simplified Grammar.

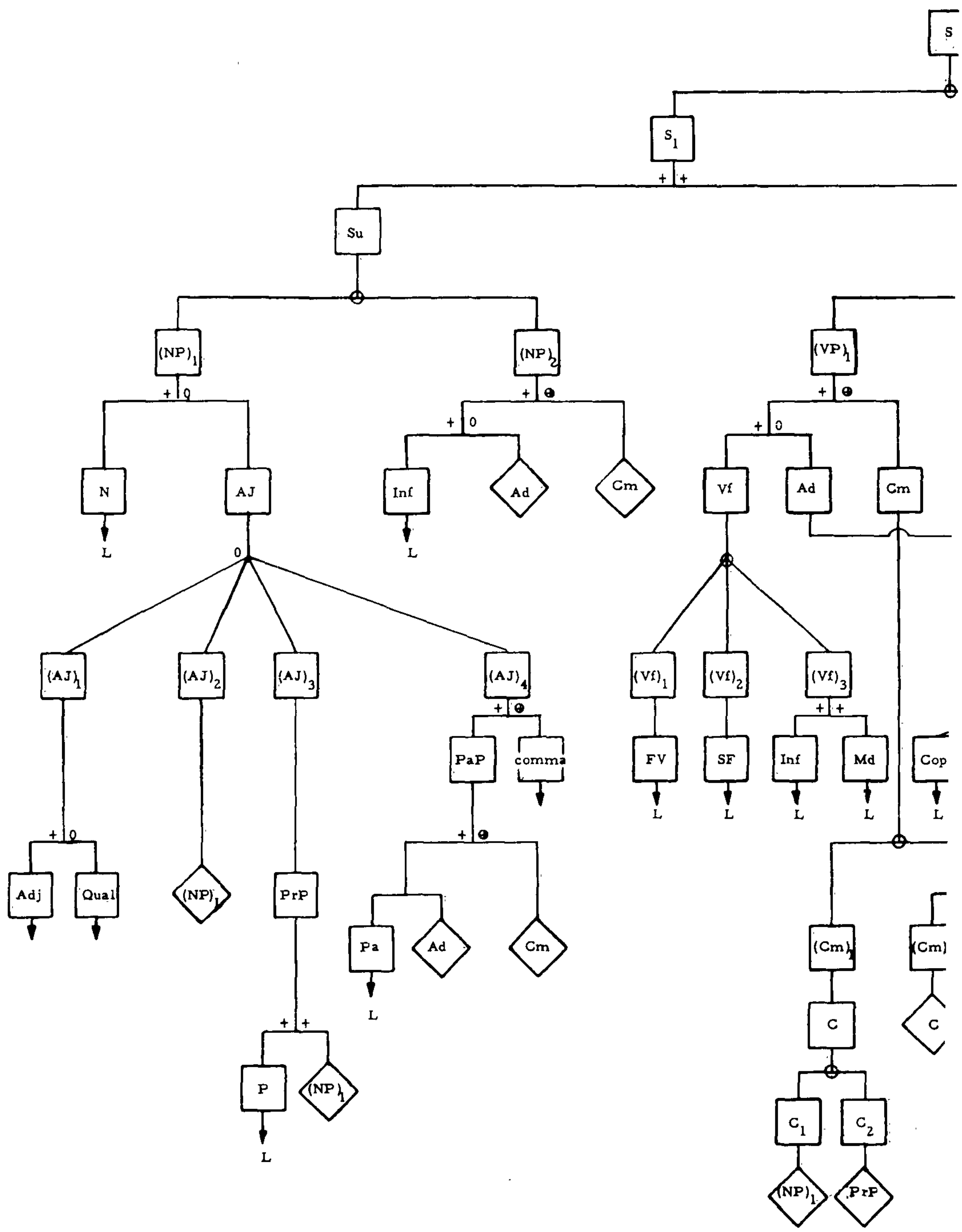
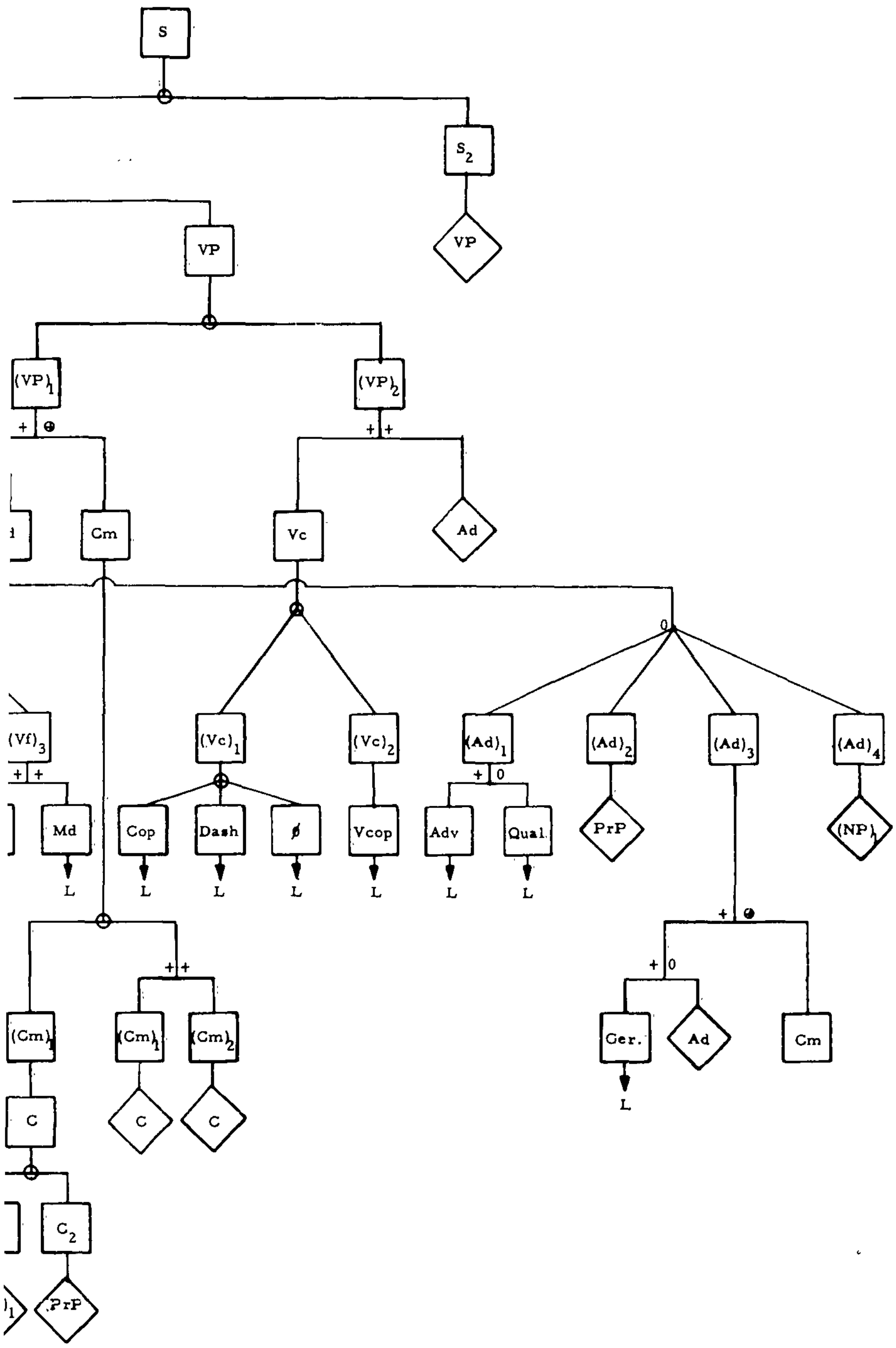


Figure 2.2-2 Diagram of a Conventional Russia



nal Russian Grammar.

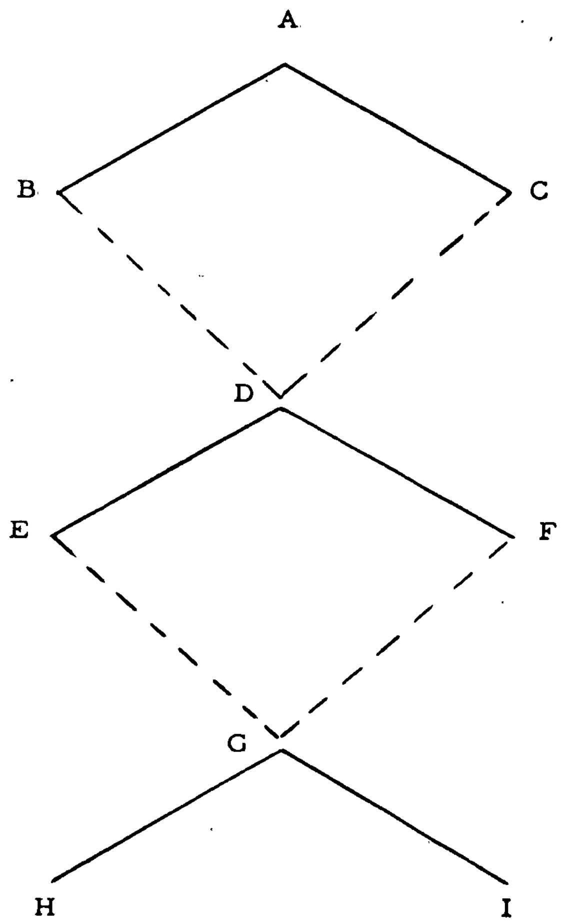
2

Ambiguous sentences or ambiguities associated with partially analyzed sentences could, in principle, be represented by a set of trees, one tree for each interpretation. If this were done, it would usually be observed that the various trees in the set had large parts in common. In Russian, for example, a sentence with a relative clause might be unambiguous except with respect to the noun modified by the relative clause. Then, instead of having several complete trees, we could describe the sentence by a main tree representing the sentence minus the relative clause, and a second tree representing the relative clause, with an appropriate indication that the relative clause tree can be attached to the main tree at any one of several points. If the relative clause contained an ambiguity within it, it in turn could usually be represented as a main tree and a subtree without altering the remaining part of the representation.

The space-saving achieved by the type of representation described above is quite impressive. If a sentence contained ten independent ambiguities, more than 1000 complete trees would be required, while the main tree-subtree method would require only 10 partial trees. Moreover, the space occupied by the 10 partial trees plus the appropriate indicators would not be much greater than the space occupied by a single one of the 1000 complete trees.

Even more important than the space-saving is the fact that in addition to representing the total ambiguity economically, the main-tree-subtree method yields a manipulatable representation of what might be called the structure of the ambiguity. This manipulatability is essential for practical machine recognition of sentence structure. During the process of recognition, temporary ambiguities associated with partial analysis will continually arise and then be resolved as the analysis proceeds. It is obvious that a representation in the form of a set of complete trees would be hopelessly inadequate to handle this situation.

In the example of multiple ambiguity used above, the individual ambiguities were independent. The situation is more complicated in the case of dependent or conditional multiple ambiguities. Consider the following representation:



There are two ambiguities here: node D can be attached either to B or C, and node G can be attached either to E or F. The ambiguities are called independent if D can be attached to B or C independently of whether G is attached to E or F, and vice versa. They are called dependent if the above representation is accompanied by conditions such as: if D is attached to B then G must be attached to E and if D is attached to C then G must be attached to F, or, possibly, if G is attached to E then D may be attached to either B or C but if G is attached to F, then D must be attached to B. Dependent ambiguities are very common in natural language, and are usually much more complicated than this simple example.

A housekeeping format has been developed for handling the information associated with a node in an ambiguous tree (see Figure 2.2-3). This information consists essentially of the possible nodes above, the possible branches below, possible candidates for each branch below, and the dependencies or conditions which exist between nodes above and branches below or between different branches below. The types of conditions are indicated in Figure 2.2-3 by the symbols $\pm$, α, β, γ, with the following interpretation:

- $+$ indicates required presence (occurrence) of some specified constituent or grammatical feature.
- $-$ indicates required absence (nonoccurrence) of some specified constituent or grammatical feature.
- α indicates an agreement or concordance condition
- β indicates a government condition
- γ indicates a word-order condition

The section in Figure 2.2-3 labeled Address (absolute or floating) and Entry Number in State List are included for future use in connection with part (2) of the linguistic statement. However, they cannot be discussed in detail here because part (2) has not yet been elaborated in sufficient detail.

2.2.2 Search Strategy

2.2.1 <u>Introduction.</u> By search strategy we mean the types and the

A	Name	Address (if floating)	Entry no. in State List

B	No. of Nodes Above	1	Name	Address (abs. or floating)	No. of Cond.	a	Address (node and int.)	± αβγ	± αβγ
						b	machine address of next cond. info.		
		2	machine address of next node above info.						

C	No. of Oblig. br.	1	Name	Address	No. of Cond.	a	Address (node and int.)	± αβγ	± αβγ	No. of Cand.
						b	machine address of next cond. info.			
		2	machine address of next oblig. br. info.							

D	No. of Op. br.	1	Name	Address	No. of Cond.	a	Address (node and int.)	± αβγ	± αβγ	No. of Cand.
						b	machine address of next cond. info.			
		2	machine address of next op. br. info.							

Figure 2.2-3 Format for the Recording of

γ

γ	No. of Cand.	1	Name	Address	No. of Cond.	a	Address (node and int.)	± αβγ	± αβγ
						b	machine address of next cond. info.		
		2	machine address of next cand. info.						

γ	No. of Cand.	1	Name	Address	No. of Cond.	a	Address (node and int.)	± αβγ	± αβγ
						b	machine address of next cond. info.		
		2	machine address of next cand. info.						

rding of Node Information.

2

sequencing of the operations, manipulations, and intermediate representations used in the process of machine recognition of sentence structure. If natural language sentences were unique syntactically (i.e., if any given sentence contained sufficient grammatical constraints to permit the assignment of only one structural description), search strategy would be concerned mainly with economy. There would probably be many adequate search strategies for any given language and perhaps several general strategies applicable to any language. Even under such relatively simple conditions, search strategies would not be simple, and their experimental testing and evaluation, and theoretical investigations of their properties, would be an important area of MT research. The importance of such research becomes all the more apparent when one considers the actual properties of natural language sentences and the present state of linguistic theory applicable to language data processing.

The essential features are:

1. Syntactic ambiguity is a characteristic property of natural language sentences. There are many types of ambiguity, and in any particular area of language data processing one type may present more of a problem than some other type. The degree of ambiguity varies from sentence to sentence, and some sentences are syntactically unique. However, it is certainly not the case that syntactic uniqueness is characteristic of natural language and that syntactic ambiguity is a rare, accidental, or abnormal phenomenon.
2. Theories, models, concepts, notions, etc., of sentence structure come in all sizes, shapes, and colors. There are many competing approaches (transformational, phrase structures of various sorts; dependency, etc.)

which differ widely among themselves not only in specific details, but in methodology as well, and have very little common ground on such questions as the scope or domain of linguistic description and methods for testing and evaluating proposed theories (compare, for example, the situation in physics or chemistry where questions regarding the domain of the science and methods for evaluation of theories were effectively settled long ago and rarely occupy the attention of the working scientist).

3. Existing grammars (i.e., application of a more or less explicit theory, model, or procedure to the detailed description of particular languages) are very sketchy and incomplete with respect to the demands of automatic analysis of random text, and will continue to be so for a long time to come. Even small gaps in the grammar can lead to large gaps in the analysis of random text. This is clear in the analogous case of an automatic dictionary. If the dictionary has only a 5% gap, i.e., 95% of the words in a text are in the dictionary, then, ignoring nonuniformities in distribution, if the text consisted of sentences 20 words long, only 35% of the sentences would be free of missing words. Thus, a 5% gap in the dictionary would lead to a 65% gap in analyzable sentences. Actually, sentences with missing words can be partially analyzed if there is sufficient syntactic redundancy, and in the case of highly inflected languages such as Russian, part of the missing grammatical information can sometimes be supplied by morphological analysis. However, such "emergency"

procedures, while useful, are of limited effectiveness. The fact remains that a relatively high degree of dictionary coverage (probably at least 95%) is required as a basis for syntactic analysis at the sentence level. Incompleteness of grammatical description is to a certain extent analogous to gaps in dictionary coverage in that significant information is missing.

4. During the analysis of a sentence one must operate with intermediate representations which have no explicit counterparts in the traditional structural descriptions of sentences. This is not a peculiarity of MT, but is typical of the use of machines to perform tasks previously done by humans. Convenient machine procedures may not parallel the customary human procedures. Even if the machine procedure is based on the human procedure as a model, it is frequently found that some of the corresponding human operations are performed "unconsciously" and must be made "conscious" or explicit for the first time in order to simulate them on a machine. In any event, whether a search strategy is intended as a model of the human procedure (a model of the hearer) or is based on other considerations, it will require study of the properties of structures and ambiguities associated with partially analyzed sentences.

All four of the above items involve syntactic ambiguity. In item 1, it enters as an inherent characteristic of natural language. In items 2 and 3, it results from incompleteness of available grammars and linguistic theories. In item 4, it is associated with representations of partially analyzed sentences. Each of these sources of ambiguity can exist independently of the others. Some linguists might object to item 1 and argue that syntactic ambiguity is not an inherent characteristic of natural language but only seems so because of the inadequacies of available grammars. This is an important theoretical question but has very little relevance to the present discussion of search strategies because the only immediate consequence of one or another answer to the questions would be a relabeling of the source of a class of ambiguities. The ambiguities would remain. Even if items 1, 2, and 3 were eliminated in some way as sources of syntactic ambiguity, item 4 would remain as a source of ambiguity (or perhaps it would then be called pseudo-ambiguity). In any event, it is clear that one necessary component of a search strategy is the capacity to handle syntactic ambiguity.

The position of automatic sentence structure determinations and associated search strategies within the framework of linguistic theory is an open question at present. Many (perhaps most) linguists consider it to be either completely outside the domain of linguistics (a programming problem) or "merely" an interesting application of linguistic theories and procedures; i.e., if not for the demands of MT, IR and other areas of language data processing, it would be of very little or no linguistic interest. On the other hand, Chomsky considers the domain of linguistic theory to include automatic sentence structure determination as an integral part, without reference to practical applications. So far, the only theoretical investigation in this area is Matthew's analysis-by-synthesis recognition procedure.

2.2.2.2 Inadequacy of Current Classification of Search Strategies. There are about as many search strategies at present as there are MT projects. These strategies are variously described as left-to-right, right-to-left, single-pass, multipass, iterative, top-to-bottom, bottom-to-top, predictive, fulcrum, etc. However, none of these names describes a search strategy, but only one or another aspect of a search strategy. Left-to-right and right-to-left describe the order in which an input string is scanned. Single-pass, multipass, and iterative describe the number of times an input string is scanned. Top-to-bottom and bottom-to-top describe the order in which the various levels of the structural representation are determined. In top-to-bottom, the major components of the sentence (subject, main verb, object) are identified before the minor components (adjectivals, adverbials). In bottom-to-top, fragments such as noun phrases, prepositional phrases, participial phrases, etc., are identified before their function at the sentence level is established. Since these three aspects of a search strategy (order of scanning the input string, number of scans, order of determination of the various levels of the structural description) are at present considered more or less independent of one another, describing a search strategy in terms of only one of these aspects tells very little about it, especially since most of the above-mentioned procedures are idealizations unattainable in practice.

2.2.2.3 <u>Preliminary Attempt to Establish a Typology of Search Strategies</u>. Since, as indicated elsewhere, the currently accepted classification of search strategy is inadequate, the first step in search strategy research must be an attempt to establish a meaningful typology of search strategies based on essential rather than superficial features. Then, the various types and features can be compared and evaluated experimentally with regard to areas of effectiveness, economy, generality, etc. These experiments are simultaneously a test of the initial typology and selection of essential features and they will provide motivation for modifications.

The classification presented below is intended to serve only as a starting point for research and is expected to undergo more or less radical revision and extension as the research progresses.

The three main criteria are:

1. number of passes required for a <u>single interpretation</u> of a sentence

2. organization of the passes
3. order of construction of structural representation

Under point 1, we will have either one pass or many passes, i.e., only two subclasses. The inclusion of <u>single interpretation</u> will avoid confusion of strategies requiring more than one pass for unambiguous sentences with strategies which require only one pass for unambiguous sentences and yield one interpretation per pass for ambiguous sentences. Under point 2, we will have:

a. an ordered set of passes which are passed through only once for a single interpretation of a sentence.

b. an ordered set of passes which are passed through more than once for a single interpretation of a sentence.

c. an unordered set of passes or an ordered set with loops, the actual sequence of passes for a given sentence being a function of the structure of that particular sentence.

Under point 3, we will have:

a. definite sequence of levels with the same order for every sentence.

b. definite sequence of levels but order is a function of sentence type.

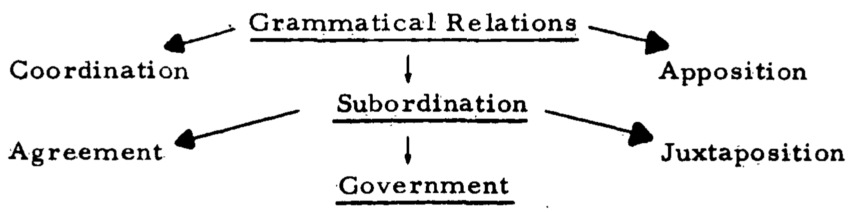
c. simultaneous (several levels are constructed simultaneously).

Summarizing and assigning tentative names to some of the types, we have the following:

Main Class	Subclass	Tentative Name
1. Number of passes	a. one	single pass
	b. more than one	plural pass
2. Organization of the passes	a. ordered set passed through only once	multipass
	b. ordered set passed through more than once	iterative
	c. unordered set or ordered but with loops activated by sentence under analysis.	recursive
3. Order of construction of structural representation	a. definite sequence, same order for all sequences	linear and ordered
	b. definite sequence, order a function of sentence type	linear and unordered
	c. several levels constructed simultaneously	simultaneous

2.2.3 Government Information

2.2.3.1 General Discussion The Russian Master Dictionary extant at IBM Research contains a complete morphological specification of Russian insofar as inflection is concerned and an operationally adequate specification of derivation within Russian. This information represents very basic data essential to subsequent automatic syntactic analysis. The next specification required is a complete or operationally adequate specification of Russian syntax. Whether an operationally adequate syntactic specification which is less than complete is possible will be determined by the development of the evolutionary program in MT. Even an operationally adequate syntactic specification (or perhaps particularly an operationally adequate syntactic specification) requires a very careful formulation of rules on the basis of available syntactic data in order to avoid the pitfalls of ad hoc rules. Only close and exhaustive study of the data can achieve this result. These data are all the syntactic relationships within and among sentences of the Russian language. It is practically impossible to divide this universe of investigation into smaller and more manageable units because of the close and critical nature of the interrelationship of the parts, but one body of data can be gathered practically independently of the study involved. This body of data is termed government in traditional linguistic parlance, and in the traditional view of grammar it occupies the following position:



All of the above grammatical relations must be explicated in terms of a much more powerful theory of grammar than that implied by the labels above. In the area of government, for example, the grammatical essence of the entire sentence

ON BORETS4 ZA SVOBODU SLOVA

(HE FIGHTS FOR THE FREEDOM OF SPEECH)

would have to be specified by an exact set of rules linking the structure of this sentence with the structures of all other sentences in the language. In this sentence the verb BORETS4 (FIGHTS) is said to govern the prepositional phrase ZA SVOBODU (FOR THE FREEDOM). In addition, the above sentence would have to be rigorously and exactly related to the following sentence:

MY SLYWALI O EGO BOR6BE ZA SVOBODU SLOVA

(WE HEARD OF HIS FIGHT FOR THE FREEDOM OF SPEECH)

where the entire original sentence has been transformed into a noun phrase with a related verbal noun BOR6BE (FIGHT) exhibiting the same government characteristics as its verb BORETS4. Obviously, such formulations are dependent on intensive study of extensive data.

These two tasks -- study of data and the compilation of data -- are certainly closely related. It is easy to claim that data gathering has enough of a priority over data investigation to justify pursuance of the former at least simultaneously with the latter. The problem is to determine whether study and compilation are so closely related that they must be carried out simultaneously or that the priority of study is mandatory because of the ill-defined nature of the data. In other words, can the domain of the data be delimited either accurately or broadly enough so that an extensive compilation program can be instituted with reasonable assurance that whole segments of the data need not be reviewed?

Study of this problem at IBM Research indicates that the data base can be made sufficiently broad enough within reason to permit wholesale compilation of grammatical data which can be loosely termed government. The succeeding paragraphs will be devoted to a presentation and discussion of what material may be subsumed under the label of government and what other grammatical data may be implied by, and included in, a study of government.

The discussion of government will be outlined in more or less the traditional terms of Soviet Russian grammars which were drawn upon heavily in the initial studies conducted on the data base. Features of government are characteristic of verbs, nouns, and adjectives. For all three parts of speech, government may be subdivided as follows:

- Government
  - Direct
    - Strong
    - Weak
  - Prepositional
    - Strong
    - Weak

These various aspects or putative aspects of government as traditionally conceived may be exemplified by verbal government in the following single sentence:

XITROST6H ON ZAXVATIL RYBU ZA JABRY V PRUDU.
(HE SEIZED THE FISH BY THE GILLS WITH CUNNING IN THE POND).
where strong direct government is represented by the noun RYBU (FISH) in the accusative case, where weak direct government is represented by the noun XITROST6H (CUNNING) in the instrumental case, where strong prepositional government is represented by the prepositional phrase ZA JABRY (BY THE GILLS), and where weak prepositional government is represented by the prepositional phrase V PRUDU (IN THE POND). In the distinctions drawn above the label "strong" seems to denote a grammatical element of central importance to the verb of the sentence while the label "weak" seems to denote a grammatical element that is marginal, casual or adventitious in the sentence. Admittedly, however, these labels have been intuitively applied; and it remains for rigorous grammatical formulation to specify the real basis or bases for the

distinction. The following paragraphs will list further examples of these types of government for verbs, nouns, and adjectives.

2.2.3.2 Verbal Government. By far the richest and undoubtedly the basic government patterns are to be found among the verbs. Once real progress has been made in the study of verbal government in terms of entire sentences and their interrelationships in Russian, then the way will be cleared to a clarification of nominal and adjectival government. Additional examples of verbal government appear below with the governed structures underlined in both Russian and English. These examples should serve to illustrate amply the varied structures that may be governed and the amount of study necessary to fit them into a theory of Russian grammar.

ON ZAVISIT OT MEN4 (prep. phrase) POMO56H (noun-instr.).
(HE DEPENDS ON ME FOR HELP.)
ONI VYPUSTILI EGO (noun-accus.) KANDIDATOM (noun-instr.)
(THEY PROMOTED HIM AS A CANDIDATE.)
ON VYRASTAET UMOM (Noun-instr.).
(HE IS MATURING IN MIND.)

KOMMUNIZM GROZIT NAM (noun-dative).
(COMMUNISM THREATENS US.)

DETI DVIGAHT RUKAMI (noun-instr.).
(THE CHILDREN ARE MOVING THEIR ARMS.)

ON DOSTAL EMU (noun-dative) DO PLECA (prep. phrase).
(HE REACHED UP TO HIS SHOULDER.)

3TA DOROGA DOVODIT DO GORODA (prep. phrase) LESOM (noun-instr.).
(THIS ROAD LEADS AS FAR AS THE TOWN BY WAY OF THE FOREST).

4 ZVONH MONETAMI (noun-instr.).

(I AM JINGLING COINS.)

ON ZALIVALS4 SOLOV6EM (noun-instr.).

(HE BURST INTO SONG LIKE A NIGHTINGALE.)

GLUBINA OZERA ZAMER4ETS4 MES4QAMI (noun-instr.).

(THE DEPTH OF THE LAKE IS MEASURED MONTH IN AND MONTH OUT.)

DVER6 ZAKRYVALAS6 PERVOI (adj.-instr.)

(THE DOOR WAS CLOSED FIRST.)

SOBAKA ZAPUTALAS6 NOGAMI (noun-instr.) V SET6 (prep. phrase).

(THE DOG GOT HIS FEET CAUGHT IN THE NET.)

In the examples above, the frequent and varied use of the instrumental case is to be noted. The frequency and variety of instrumental usages poses a severe problem not only for subsequent study of this case but also for gathering sufficient information in a compilation effort of real magnitude. Solution of this problem will necessitate at least a semantic classification of instrumental types with an accompanying test set of nouns to be put into the instrumental singular or plural and matched against each verb. The last four sentences above underscore another extremely important study area and data-gathering area among the verbs. This is the area defined by the reflexive verbs in Russian. This is an important area of investigation, for reflexive verbs are related in fundamental and varied ways to their corresponding non-reflexive forms. In addition, at times these reflexive verbs display government patterns seemingly unpredictable in terms of those belonging to their non-reflexive counterparts. Because of its fundamental importance and complexity, the subject of reflexive verbs will be discussed in some detail at the end of this section on verbal government.

Closely connected with direct government is the usage of certain verbs with infinitives and CTO-clauses (that-clauses). In the case of infinitives the infinitival phrase seems to be the transform of another sentence. For example, in the sentence:

ON GROZITS4 PRI1TI V GOSTI.

(HE THREATENS TO COME VISITING.)

the infinitival phrase PRI1TI V GOST1 would seem to represent a transform of the sentence:

ON PRIDET V GOSTI.

(HE WILL COME VISITING.)

Note that in this case the transformed sentence shares the subject with the main verb. Likewise, in the sentence:

ONA SKAZALA MNE SEST6 NA STUL.

(SHE TOLD ME TO SIT IN THE CHAIR.)

the infinitival phrase SEST6 NA STUL would seem to be a transform of the following sentence:

4 S4DU NA STUL.

(I WILL SIT IN THE CHAIR.)

Note that in this latter case the transformed sentence does not share the subject with the principal verb.

CTO-clauses, on the other hand, are transforms of complete sentences in which the entire sentence becomes the direct object of the principal verb and the transformed sentence undergoes no permutation. For example, in the sentence

ONI SKAZALI, CTO ON BUDET BOLEN

(THEY SAID THAT HE WOULD BE SICK.)

the stretch CTO ON BUDET BOLEN is the direct object of the verb SKAZALI and consists of the subordinating conjunction CTO plus the independent sentence,

ON BUDET BOLEN

(HE WILL BE SICK.).

It is to be noted that the original sentence during transformation has not changed at all whereas the corresponding English sentence has undergone a change in accordance with the sequence-of-tense rules operative in English. Another example is to be found in the following sentence:

ON DOKAJET, CTO 4 BYL VINOVAT.

(HE WILL PROVE THAT I WAS GUILTY.)

where the subordinate clause introduced by CTO is again the direct object of the main verb and where the sentence of the subordinate clause has been transformed unchanged.

Obviously, these data must be encompassed by a compilation effort and are, in fact, much more easily accounted for than manifestations of weak government as outlined under direct and prepositional government above.

2.2.3.3 Reflexive Verbs. As previously stated, the importance of reflexive verbs in any data-gathering operation such as that in preparation at IBM Research is not to be overlooked. The category of reflexivity in the Russian language has long commanded the attention of grammarians and linguists. Reflexivity, as a category involving the relationships between the subject and object of an action, is properly a category of voice. Undoubtedly, the most succinct characterization of the semantic content of this grammatical category has been made by Roman Jakobson who dichotomizes the category into reflexive versus non-reflexive and states that the reflexive restricts participation in the narrated event, while the non-reflexive says nothing about restriction of participation. If the non-reflexive verb is transitive with two participants in the narrated event, one of them is eliminated by the reflexive. Thus, the sentence

SOLDATY ZA5I5AHT KREPOST6

(THE SOLDIERS DEFEND THE FORTRESS.)

yields, when the reflexive particle is added, either

SOLDATY ZA5I5AHTS4

(THE SOLDIERS DEFEND THEMSELVES.)

See Jakobson, Roman: Shifters, Verbal Categories, and the Russian Verb. 1957 Publication of Russian Language Project, Harvard University.

or

KREPOST6 ZA5I5AETS4

(THE FORTRESS IS BEING DEFENDED.)

If the non-reflexive verb is intransitive, the grammatical subject may be excluded or the sphere of action may be restricted. For example, the sentence

ON XOROWO JIVET

(HE LIVES WELL.)

may yield the impersonal reflexive

XOROWO JIVETS4

(ONE LIVES WELL.)

And, furthermore, the sentence

ON STUCIT

(HE IS POUNDING.)

becomes particularized upon addition of the reflexive particle to give

ON STUCITS4

(HE IS KNOCKING TO BE ADMITTED.)

In any rigorous formulation of Russian grammar, envisioned elsewhere in this report and expected to lead to a truly explanatory grammar of Russian, the practically all-pervasive category of reflexivity will play a very important role and must be exhaustively investigated. It seems expedient at this stage of the work merely to outline the broad semantic usages of the reflexive voice so that all verbs encountered can be preliminarily classified for subsequent study. Accordingly, the following outline of reflexive usages is presented as a discussion guide to be followed by grammarians collecting data for grammatical study and improvement of Russian-English machine translation.

a. Reflexives Proper

General: The agent of the action is at the same time its object. The action is not directed to an outside object but rather "returns" to the agent. The particle "-S4" is generally equal to the reflexive pronoun SEB4.

The meaning of reflexivity is clearest when the verb is used in its non-reflexive form with SEB4 as its direct object. Within present norms of the Russian language such constructions are not always interchangeable. However, their similarity is strongly felt. Note the following examples:

ON PRIKRYL SEB4 ODE4LOM and ON PRIKRYLS4 ODE4LOM
(HE COVERED HIMSELF WITH A BLANKET.)

JEN5INY CASTO RASSTRAIVAHT SEB4 PONAPRASNU and
JENSINY CASTO RASSTRAIVAHTS4 PONAPRASNU
(WOMEN OFTEN BECOME DISTRESSED UNNECESSARILY).

It may, in fact, be argued that in the last example the second sentence has strong overtones of what will be discussed later as generally reflexive meaning.

Within the reflexive forms of the verb the meaning is clearest with verbs denoting the action directed at the body of the agent who must be described by an animate noun.

UMYVAT6S4	(WASH ONESELF)
PUDRIT6S4	(POWDER ONESELF)
KUTAT6S4	(BUNDLE ONESELF)

are some of the verbs which fall into this category.

The above meaning is somewhat weakened when the action is directed at the <u>person</u> of the agent:

ZA5I5AT6S4	(DEFEND ONESELF)
OPRAVDYVAT6S4	(JUSTIFY ONESELF)
SDERJIVAT6S4	(RESTRAIN ONESELF)

The majority of the properly reflexive verbs, especially in their non-reflexive forms fall into the category of "concrete action" verbs as described in the traditional grammar.*

Some of the "non-concrete" action verbs, although close to the meaning of reflexives should be classified under the category of generally reflexive verbs. Thus, in the case of such verbs as

OSVEJIT6S4	(FRESHEN UP),
UTOMIT6S4	(GROW TIRED),
POZABAVIT6S4	(AMUSE ONESELF)

and some others, it is really difficult to argue that they can be replaced by their non-reflexive form and SEB4, since what is implied is a transition from one state to another and not any specific action. Generally, the distinction here may be drawn between perfective and imperfective forms of the verbs. While the perfectives may be called reflexives in some marginal cases, the imperfective forms should, it seems, be classed as generally reflexive.

Among properly reflexive verbs, perfective forms should be translated actively irrespective of their subject. Imperfective verbs appearing with animate subjects should be translated actively as well. If, however, they combine with an inanimate subject, they become passively reflexive and should be translated passively.

b. Reciprocally Reflexive Verbs

General: This group unites verbs which describe an action taking place between two or more agents or groups of agents. The meaning of the particle "-S4" is equivalent to:

---

* In the present analysis, the distinction between "concrete" and "non-concrete" action verbs is decided solely by the ability to combine with a group of adverbs of degree: OCEN6 (VERY), CUT6-CUT6 (A LITTLE), SLIWKOM (EXCESSIVELY) and some others.

DRUG DRUGA or ODIN DRUGOGO (ONE ANOTHER),

e.g., SOLDATY VSTRECAHTS4 = SOLDATY VSTRECAHT ODIN DRUGOGO

In their non-reflexive forms, where these are possible*, the above verbs require an object which is capable of the same action as the subject:

DEVUWKA QELUET MAT6 (THE GIRL KISSES HER MOTHER)

where the reverse is also possible.

MAT6 QEWET DEVUWKU (THE MOTHER KISSES THE GIRL).

The meaning of a reciprocal action is clearest when the agent of the action appearing as the subject is grammatically connected to the other agent by means of the preposition "S" governing the instrumental:

DEVUWKA QELUETS4 S MATER6H

(THE GIRL AND HER MOTHER ARE KISSING ONE ANOTHER.)

SOLDATY VSTRETILIS6 S LETCIKAMI

(SOLDIERS MET WITH THE PILOTS.).

It should be noted that while this construction is optional when at least one of the agents is in the plural, it is obligatory when both are in the singular.

The above construction with "S" may with some verbs introduce the meaning of a mutual action:

OB7EDIN4T6S4	(UNITE)
WEPTAT6S4	(WHISPER TO ONE ANOTHER)
SSORIT6S4	(QUARREL)

and some others.

---

* Note that the relationship between the reflexive and non-reflexive form is frequently complicated (SOVETOVAT6S4 - TO CONSULT, PEREPISYVAT6S4 - TO CORRESPOND WITH SOMEONE) or lost (RASXODIT6S4 - TO DISPERSE, SOREVNOVAT6S4 - TO COMPETE).

Further, there are some non-reflexive verbs with reciprocal meaning: BESEDOVAT6 (TO CONVERSE), SPORIT6 - (TO ARGUE) DRUJIT6 - (TO BE FRIENDS WITH SOMEONE).

At present, however, this distinction is not very significant for machine translation.

A variant of the preceding is represented by some of the intransitive verbs of motion which can combine with the prefixes RAZ- and S- and the particle "-S4."

SXODIT6S4 (COME TOGETHER)

RAZBRESTIS6 (WANDER APART)

Some of the other reflexive verbs formed by prefixation are discussed elsewhere (see (f) below).

As is true of properly reflexive verbs, most reciprocally reflexive verbs are used with animate subjects. Passivity is not apparent in cases where the inanimate subject is found. It should be noted also that the meaning of many of imperfective reciprocal verbs tends to approach that of active objectless verbs (see (d) below).

In general, the translation should be active for both perfective and imperfective forms.

Some problems arise in regard to generalized statements like

TAKIE LHDI REDKO VSTRECAHTS4

where the ambiguity between

SUCH PEOPLE SELDOM MEET

and

SUCH PEOPLE ARE SELDOM MET

would have to be resolved by context and thorough study of the problem. The same is true of the meaning of instrumental constructions which would have to be studied in conjunction with currently proposed research into the grammar of Russian.

c. Passive Reflexives

General: Passive reflexives are found in constructions where the grammatical subject is the object of the action and the agent of the same action is expressed by the grammatical object in the instrumental case.

KUXARKA GOTOVIT OBED -- OBED GOTOVITS4 KUXARK01

(THE COOK PREPARES THE DINNER -- THE DINNER IS BEING PREPARED BY THE COOK.).

The passive meaning is clearest when the reflexive verb is in its imperfective form and is initially derived from a directly-transitive verb. Of the obliquely transitive verbs only those that govern the instrumental case

KOMANDOVAT6	(TO COMMAND);
UPRAVL4T6	(TO GOVERN, DIRECT).
DIRIJIROVAT6	(TO CONDUCT)

and others can function in a similar capacity.

Subject to restrictions to be worked out in the planned intensive study of Russian grammar, such constructions, especially when they appear with inanimate subjects, are to be translated passively.

Obliquely-transitive imperfective verbs with the exception of those governing through the instrumental case, and all transitive perfective verbs, especially when they appear with an inanimate subject, are unlikely to have reflexive forms of the verb used to express passivity and would instead be used with passive participial constructions.

While the perfective passive reflexives are a moot issue, the only recognized exception to the restrictions formulated above are cases of transition from one state to another and are expressed with the additional qualification that such transition occurs "independent of the will of the agent":

KOMNATA OSVETILAS6 TAINSTVENNYM SVETOM

(THE ROOM BECAME ILLUMINATED BY A MYSTERIOUS LIGHT

POL4 POKRYLIS6 SNEGOM

(FIELDS BECAME COVERED WITH SNOW)

LES OKUTALS4 TUMANOM
(THE FOREST BECAME ENVELOPED BY FOG.).

It should be emphasized again that should even such a framework be accepted any final formulation would have to resort to a much more rigid definition of criteria for judgments.

Any mixed case of criteria stated above should be decided in favor of the imperfective or accusative. Thus the verbs which are both perfective and imperfective should be treated as imperfective and verbs which have "double" government, if they have the accusative government as one of the possibilities, -- as directly transitive.

In many instances, passive reflexives, when used without the grammatical object in the instrumental, tend to approach the meaning of generally reflexive verbs. Thus:

NOVYE GRANATY VZRYVAHTS4 MEDLENNO OSOBYM DETONATOROM

(NEW GRENADES ARE EXPLODED SLOWLY BY A SPECIAL DETONATOR.).

However, this sentence without the indirect object "OSOBYM DETONATOROM" (SPECIAL DETONATOR) becomes ambiguous and suggests the range of changes possible in transition from the passively reflexive to generally reflexive:

NOVYE GRANATY VZRYVAHTS4 MEDLENNO

a. NEW GRENADES EXPLODE SLOWLY

b. NEW GRENADES ARE EXPLODED SLOWLY

The clarity of passive meaning depends on the subject; and hence, as a rule, passive meanings are better expressed when the subject is inanimate because only all animate being can be both a subject and the object of an action. Hence, in many instances, especially where both the subject and the object of the action are equally capable of performing it, the passive meaning is obscured and is at least stylistically impossible and, sometimes, the meaning is outright ambiguous. One example of this may be the following*: Thus if

REBENOK ODEVAETS4 N4NE1

could be understood as basically meaning that

"THE CHILD IS BEING DRESSED BY A NURSE"

and not that

"THE CHILD IS DRESSING ITSELF AS A NURSE"

the ambiguity in the sentence

SOLDAT ODEVAETS4 MONAXOM

(THE SOLDIER IS BEING DRESSED BY A MONK)

or

(SOLDIER DRESSED HIMSELF AS A MONK)

is even more apparent. If in the same constructions the verb is perfective and is used in the past tense, the passive meaning is not felt

REBENOK ODELS4 N4NE1, SOLDAT ODELS4 MONAXOM.

Finally, in the participial form the meanings are differentiated completely:

REBENOK ODEVAEMY1 N4NE1

(THE CHILD WHO IS BEING DRESSED BY A NURSE)

and

SOLDAT ODEVAH5I1S4 MONAXOM

(SOLDIER WHO DRESSES HIMSELF AS A MONK)

What needs to be pointed out then is that although there are some morphological markers for the passive meaning, there are syntactic and semantic problems which intermingle and influence passive constructions with reflexive verbs.

d. Generally Reflexive Verbs

General: This group of verbs is formed from transitive non-reflexive verbs and is characterized by the weakened tie of the action

---

* Note the variations occurring with inanimate, indirect objects.

to its object and a concentrated emphasis upon the agent of the action. The feature which distinguishes the generally reflexives is that these verbs cannot be replaced by non-reflexive forms. E.g.,

ON VOLNUET MAT6 (HE UPSETS (HIS) MOTHER)

but ON VOLNUETS4 (HE IS UPSET)

ON VOSXI5AET VSEX SVOIMI SPOSOBNOST4MI (HE DELIGHTS EVERYONE WITH HIS ABILITIES.)

by ON VOSXI5AETS4 SVOIMI SPOSOBNOST4MI (HE IS DELIGHTED BY HIS ABILITIES.).

The above definitions most clearly apply to verbs denoting "physical or mental conditions" of the agent -- in this case ANIMATE SUBJECT. However, corresponding non-reflexive forms can combine with both animate and inanimate nouns.

Hence, the passive meaning is impossible here because the agent himself is "engulfed" by the action.

The point is clear if one compares the following list:

RADOVAT6 (MAKE SOMEONE GLAD)	RADOVAT6S4 (BE GLAD)
VOLNOVAT6 (AGITATE)	VOLNOVAT6S4 (BE AGITATED, BE IN AGITATION)
VESELIT6 (CHEER)	VESELIT6S4 (ENJOY ONESELF)
UDIVL4T6 (ASTONISH)	UDIVL4T6S4 (BE ASTONISHED)
PECALIT6 (SADDEN, GRIEVE)	PECALIT6S4 (BE SAD, GRIEVE)
SERDIT6 (ANGER)	SERDIT6S4 (BE ANGRY)

VOSXI5AT6 (DELIGHT)	VOSXI5AT6S4 (BE DELIGHTED/ADMIRE)
STRAWIT6 (FRIGHTEN)	STRAWIT6S4 (BE FRIGHTENED)
BESPOKOIT6 (DISTURB)	BESPOKOIT6S4 (BE DISTURBED)
UTEWAT6 (CONSOLE)	UTEWAT6S4 (BE CONSOLED)
PODCIN4T6 (SUBORDINATE)	PODCIN4T6S4 (BE SUBORDINATED)
TOROPIT6 (HURRY)	TOROPIT6S4 (HURRY, BEING IN A HURRY)
TREVOJIT6 (UPSET, DISTURB)	TREVOJIT6S4 (BE UPSET, DISTURBED)
USPOKAIVAT6 (CALM)	USPOKAIVAT6S4 (BECOME CALM)
UDIVL4T6 (ASTONISH)	UDIVL4T6S4 (BE ASTONISHED)
ZLIT6 (ANGER)	ZLIT6S4 (BE ANGERED)
TEWIT6 (AMUSE)	TEWIT6S4 (AMUSE ONESELF)
INTERESOVAT6 (INTEREST)	INTERESOVAT6S4 (BE INTERESTED)
VOZBUJDAT6 (EXCITE)	VOZBUJDAT6S4 (BE EXCITED)
POKOR4T6 (SUBJUGATE, SUBDUE)	POKOR4T6S4 (RESIGN ONESELF/SUBMIT)
ZABAVL4T6 (ENTERTAIN)	ZABAVL4T6S4 (BE ENTERTAINED)

The meaning of verbs described above is less apparent in the group of verbs describing changes in the state or position of the agent. These changes occur in space and hence are more perceptible visually:

IZMEN4T6 (CHANGE)	IZMEN4T6S4 (CHANGE/BE CHANGED)
DVIGAT6 (MOVE)	DVIGAT6S4 (MOVE/BE MOVED)
NAGIBAT6 (BEND)	NAGIBAT6S4 (BEND/BE BENT)
RAZVIVAT6 (DEVELOP)	RAZVIVAT6S4 (BE DEVELOPED/DEVELOP)
UVELICIVAT6 (INCREASE)	UVELICIVAT6S4 (BE INCREASED)
UXUDWAT6 (WORSEN)	UXUDWAT6S4 (WORSEN, DETERIORATE)
OSTANAVLIVAT6 (STOP)	OSTANAVLIVAT6S4 (BE STOPPED/STOP)
ZAMEDL4T6 (SLOW DOWN)	ZAMEDL4T6S4 (SLOW DOWN/BE SLOWED DOWN)

Thus, the action expressed by the transitive verb reverts back to the subject which could be both animate and inanimate. The influence of the animate-inanimate subject is less significant here than the presence of an indirect object in the instrumental which seems at this time the only marker of a passive translation.

A subcategory of the above is the relatively small group of verbs denoting beginning, continuation, or termination of a phenomenon or action. This group is marked in its non-reflexive form by the ability to govern infinitives. Only inanimate subjects are possible. Passive translations are possible only when there is an instrumental construction:

REAKQI4 PREKRA5AETS4 VVEDENIEM KISLORODA
(THE REACTION IS STOPPED BY THE INTRODUCTION OF OXYGEN.).

Some of these verbs are:

NACINAT6S4	(TO START),
PRODOLJAT6S4	(TO CONTINUE),
KONCAT6S4	(TO STOP),
PREKRA5AT6S4	(TO STOP),
ZAVERWAT6S4	(TO TERMINATE).

A possible subgroup to verbs denoting change in state or position consists of verbs denoting processes and qualities of objects. These verbs are possible only with inanimate subjects:

PRODUKTIVNOST6 TRUDA POVYWAETS4
(LABOR PRODUCTIVITY IS INCREASING);

PROIZVODSTVO CUGUNA RASWIR4ETS4
(PIG IRON PRODUCTION IS EXPANDING);

PIVO PENITS4
(BEER FOAMS);

STEKLO B6ETS4
(GLASS BREAKS);

PRUJINA SJIMAETS4
(A SPRING CONTRACTS);

ODEJDA IZNAWIVAETS4
(CLOTHING WEARS OUT);

45IK VYDVIGAETS4
(THE DRAWER COMES OUT), etc.

Addition of the object in the instrumental renders these verbs passively reflexive (see (c) above).

In traditional grammars, verbs listed in this subgroup are generally singled out as expressing "qualitative-passive meaning." For machine translation purposes it is likely that the classification of generally reflexive verbs would have to be consolidated; however, for clarity of presentation and continuity the present classification seems the most efficient expedient.

Yet another variant of the previous subgroups is the verbs which are sometimes labeled as "active objectless verbs." The meaning expressed by the non-reflexive form is the same as that of the reflexive form except that the meaning of the latter is intensively manifested and is thought of as characteristic of the subject which is limited almost entirely to plant and animal nouns:

SOBAKA KUSAET LHDE1	→	SOBAKA KUSAETS4
(THE DOG BITES PEOPLE)	→	(THE DOG BITES)
KRAPIVA JJET KOJU	→	KRAPIVA JJETS4
(THE NETTLE BURNS THE SKIN)	→	(THE NETTLE BURNS)
KOROVA BODAET DETE1	→	KOROVA BODAETS4
(THE COW BUTTS CHILDREN)	→	(THE COW BUTTS)

While the verbs in the above two groups are similar in some respects, there are some significant differences especially in the relationship the two have with their respective non-reflexive forms. Whereas in the first groups the meaning of the reflexive form is different, it is not the case with verbs described in this section. Finally, the meaning of the verbs in both categories are possible only in the imperfective aspect. Verbs in the second group do not have a perfective counterpart while the other verbs can have it in some cases of reflexive forms and in nearly all cases of non-reflexive forms.

The idea of intensity of the action expressed by the reflexive form is carried on in the "obliquely reflexive" meanings. The difference between the reflexives and non-reflexives of such verbs is that in the former the agent is in fact also the indirect object of the source action which is performed for him in his interest. Most frequently the subject is animate but inanimate subjects are also possible. When an object in the instrumental occurs with an animate subject, the translation should not be passive:

MY ZAPASLIS6 DEN6GAMI
(WE HAVE PROVIDED OURSELVES WITH MONEY),

SOBIRAT6S4 V DOROGU
(PREPARE ONESELF FOR A TRIP),

RAZDOBYT6S4
(PROCURE),

STROIT6S4
(BUILD FOR ONESELF).

A so-called intensively-reflexive meaning is found in some verbs which can be formed from obliquely transitive and intransitive non-reflexive forms. Both the reflexive and non-reflexive meanings coincide lexically and can occur only with animate subjects.

STUCAT6 (S4)	(KNOCK)
QELIT6 (S4)	(AIM).
ZVONIT6 (S4)	(RING),
XVATSTAT6 (S4)	(BOAST),
GROZIT6 (S4)	(THREATEN) etc.

The action concentrates within the agent to the extent that it attracts outside attention. Addition of the prefix DO- adds the meaning of resultativeness of such an action.

Close to the meaning of verbs discussed just above is the group of intransitive verbs in -ET6 which describe visually perceptible changes in the outward appearance of the object -- most frequently color. Non-reflexive forms of such verbs usually have two meanings: to stand out by the color, or become of a certain color. Addition of the particle "S4" emphasizes precisely the first of the two meanings given. Both the lexical meanings and the subject in the reflexive and non-reflexive forms, are the same.

VDALI CERNEET KUST → VDALI CERNEETS4 KUST
(A BUSH LOOMS BLACK IN THE DISTANCE.)

This category labeled by Vinogradov as "passive manifestation of an outward characteristic" includes among others such verbs as:

BELET6S4	(APPEAR WHITE),
ZELENET6S4	(APPEAR GREEN),
TLET6S4	(SMOLDER), etc.

The last group of verbs which may be classed here are the middle-passive reflexive verbs. The reason for including it at this point lay largely in outward resemblances to some features of verbs to be discussed in the next section: reflexive verbs with impersonal meanings.

In the case of middle passive reflexive verbs, the object of an action is represented as its grammatical subject and the agent of the action is an indirect object toward which the action is directed. This indirect object occurs only with personal verbs. The relationship between the reflexive and non-reflexive constructions is seen from the following examples:

DOMA ON PRIPOMNIL VSE	→	DOMA EMU VSE PRIPOMNILOS6
(AT HOME HE RECALLED EVERYTHING)	→	(AT HOME "EVERYTHING RECALLED ITSELF TO HIM" AT HOME IT ALL CAME BACK TO HIM)

IX LIQA PREDSTAVL4HTS4 EMU PO NOCAM

(THEIR FACES "PRESENT THEMSELVES TO HIM" AT NIGHT)

in comparison to

ON PREDSTAVL4ET (SEBE) IX LIQA PO NOCAM

(HE SEES THEIR FACES AT NIGHT.)

The correspondence between the reflexive - non-reflexive forms is not always easily derived and this category merges with impersonal reflexive verbs.

e. Impersonal Reflexive Verbs

General: When a reflexive verb is used impersonally, the agent of the action appears as the object of the same action toward which this action is directed. There is an implication made in traditional grammars that such actions take place independent of the will of the agent while the non-reflexive forms express actions which are dependent on the will of the agent. The usage is limited to the 3rd person singular and to the neuter in the past.

NOC6H BOL6N01 SPAL XOROWO → NOC6H BOL6NOMU XOROWO SPALOS6

(AT NIGHT THE PATIENT SLEPT WELL) (AT NIGHT THE PATIENT COULD SLEEP WELL)

VCERA 4 RABOTAL → VCERA MNE RABOTALOS6

(YESTERDAY I WORKED) (YESTERDAY I FELT LIKE WORKING)

SOLDAT SIL6NO PROMERZ NA MOROZE → SOLDATU PROMERZLOS6 NA MOROZE

(THE SOLDIER FROZE VERY MUCH IN THE COLD)

It is necessary to single out first of all a group of impersonal reflexive verbs describing natural phenomena SMERKALOS6, STEMNELOS6, etc. which frequently do not have non-reflexive pairs.

The bulk of impersonal reflexive verbs are formed from intransitives and absolute forms of transitive verbs. The object is in the dative when it does appear.

EMU NE SIDITS4

(HE DOES NOT FEEL LIKE SITTING)...

This category of verbs and their translation can be worked out after the results of currently proposed analysis and careful study of impersonal sentences in Russian are available. Generally, the following suggestions can be advanced for practical application:

1. If the impersonal verb is preceded (or succeeded in the immediate vicinity) by a pronoun or a noun in the dative and it is followed by:

a. an adverb or adverbial modifier or an infinitive, the Russian dative should be translated as the English nominative and the verb given appropriate active translation.

VCERA NAM XOROWO SPALOS6

(YESTERDAY WE SLEPT WELL)

b. If none of the words is mentioned in "a" above, the dative should still be translated as nominative, followed by the phrase "feel like" plus the English present participle of the Russian verb.

MOLODOMU NE SIDITS4

(THE YOUNG MAN DOES NOT FEEL LIKE SITTING)

An exception to the above are some verbs which have to be translated into English by "IT...":

MNE KAJETS4, CTO ... IT SEEMS TO ME THAT ....

OKAZALOS6, CTO ... IT TURNED OUT THAT ....

2. If there is no antecedent dative, the translation could be "ONE" followed by the 3rd person singular in the appropriate tense.

ZDES6 VOL6NO DYWITS4 -- ONE BREATHES FREELY HERE

f. Compound Reflexive Forms

General: A number of reflexive verbs are derived by simultaneous addition of the particle "S4" and some prefix. One instance is the verbs discussed under (b) above. Most of the other verbs express fullness of the action with various additional meanings. The study of these verbs appears best suited in conjunction with the study of prefixation.

g. Verbs Not Used Without "-S4"

General: In this class appear verbs without non-reflexive counterparts. Only a few (approximately a thousand) verbs are used exclusively in reflexive forms. Some examples are:

SME4T6S4 -- TO LAUGH, ULYBAT6S4 -- TO SMILE

2.2.3.4 Nominal Government. Nominal government is closely linked to that of verbal government, especially in cases where the noun is a verbal noun and noun plus governed elements are a nominalization of verb plus governed elements. As in verbal government, one may distinguish in traditional terms between direct and prepositional government. Further, we may distinguish between inherent and derivational nominal government, as well as between strong and weak government. Diagrammatically these relationships can be portrayed as shown in Figure 2.2-4.

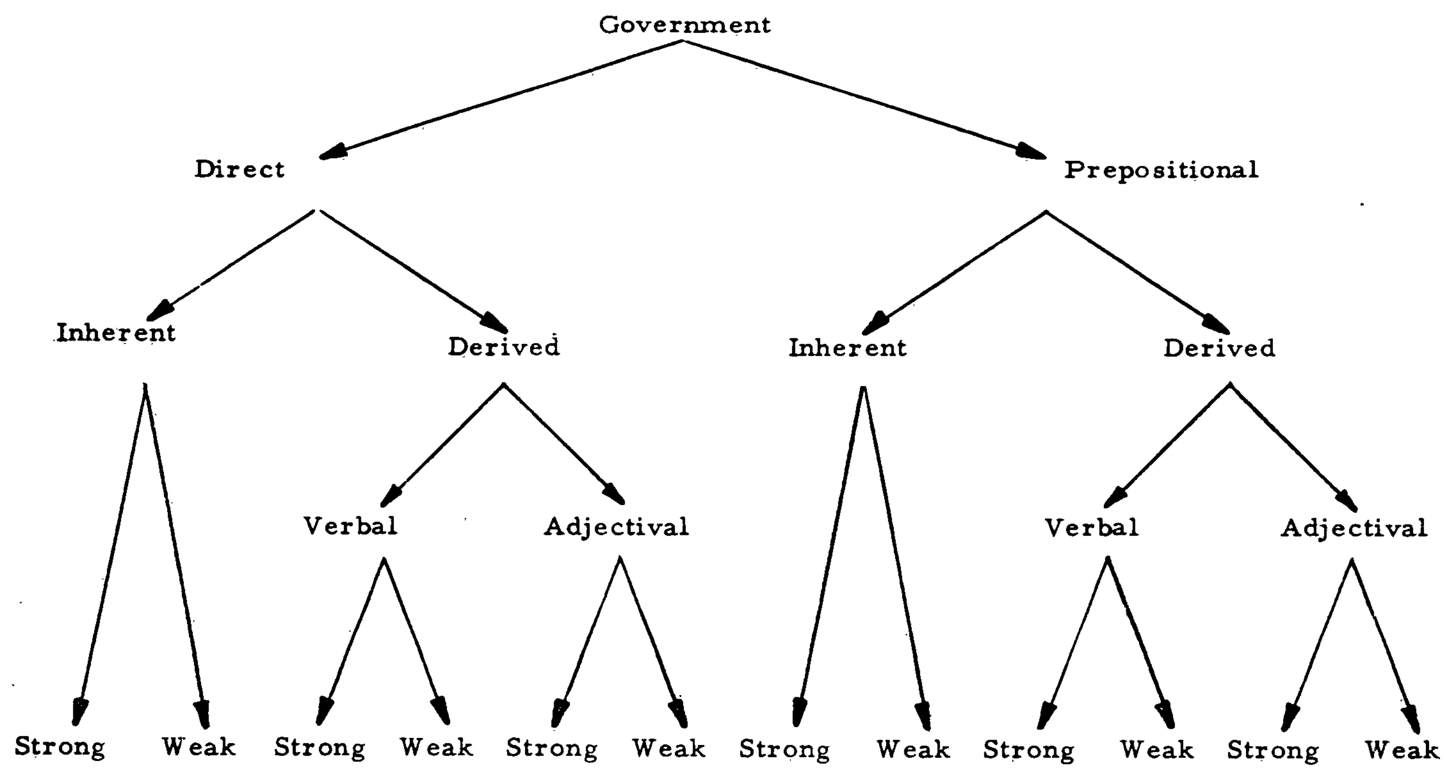


Figure 2.2-4 Diagram of Types of Nominal Government

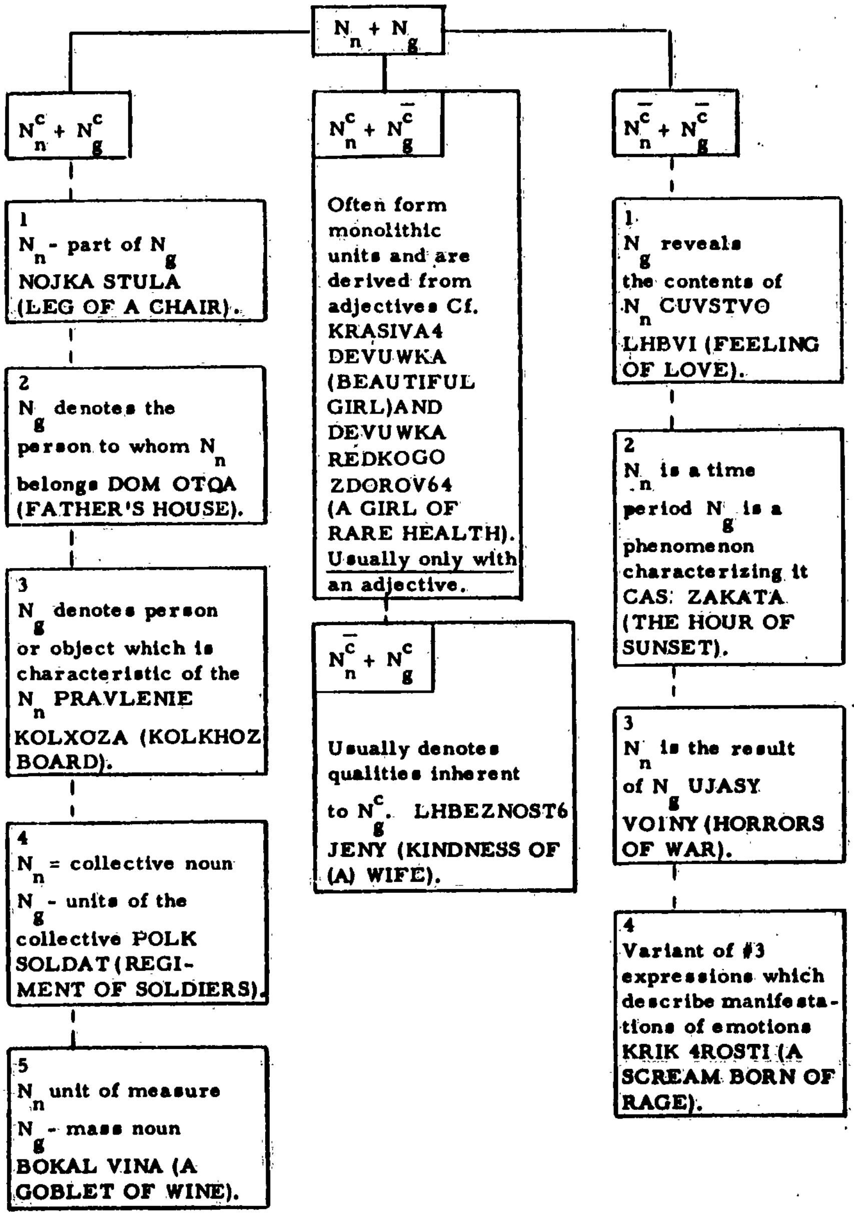
The distinction between derived and inherent government has been made to account in a rough way for the apparent priority of verbal and adjectival government. Thus, nominal expressions like VERA V PROGRESS (FAITH IN PROGRESS) and STREMLENIE K USPEXU (STRIVING FOR SUCCESS) are clearly to be derived from the verbal sentence units VERIT6 V PROGRESS (TO BELIEVE IN PROGRESS) and STREMIT6S4 K USPEXU (TO STRIVE FOR SUCCESS) whereas nominal expressions like STAKAN CAH (GLASS OF TEA), DEVUSHKA S KOSOI (GIRL WITH A BRAID), SOLDAT V WINELI (SOLDIER IN A GREAT COAT) are apparently not derived from verbal sentence units and are classed here as inherent. Nominal expressions derived from adjectives may be exemplified by NEDOVOL6STVO DRUGOM (DISSATISFACTION WITH A FRIEND) and RAZOCAROVANNOST6 V JIZNI (DISILLUSIONMENT IN LIFE), both transformed from the adjectival sentence units NEDOVOL6NYI DRUGOM (DISSATISFIED WITH A FRIEND) and RAZOCAROVANNY1 V JIZNI (DISILLUSIONED IN LIFE), but in the latter case the ultimate source is undoubtedly the verbal unit RAZOCAROVAT6S4 V JIZNI (TO BE DISILLUSIONED IN LIFE).

It should be clear from the above discussion that the traditional classification is far from being totally explanatory. For one thing it seems to be making an artificial distinction between nominals that exhibit derived government and those that exhibit inherent government. It would indubitably be much closer to the truth to say that both types of nominalizations proceed from kernel sentences, the former from sentences containing non-copulative verbs, the latter from sentences with copulative verbs. Only concentrated study can reveal the true nature of these relationships. The task at hand is to determine just what government information should be the object of data compilation.

In general, it may be stated that the so-called inherent government should not be an object of data compilation at the present time since the relationships between the governor and governed elements is too varied, too vague, and requires too much study. Their true nature would emerge only after an exhaustive study of nominalizations in Russian. Such a study will eventually be a reality, but there are presently more urgent tasks for machine translation.

One remotely possible exception is the NOUN + NOUN IN GENITIVE construction where the two nouns are not derived from a verbal syntactic unit such as ATAKA KITA1QEV (THE ATTACK OF THE CHINESE) from KITA1QY ATAKUHT (THE CHINESE ARE ATTACKING) and ZAVOEVANIE INDII (THE CONQUEST OF INDIA) from ZAVOEVAT6 INDIH (TO CONQUER INDIA). Derived nominalizations of this type could very easily be incorporated into the domain of inquiry of a data-gathering activity. So-called inherent nominalizations of this type, however, present a very broad spectrum of semantic relationships, at least, and extensive investigation alone could determine what are the basic grammatical types and whether any valid statements can be made about what nouns enter into what types. Figure 2.2-5 is a table illustrating some of the semantic varieties of this construction. As the table suggests, perhaps as much as can be known about the component nouns is that they are either concrete or non-concrete. An initial study can be carried out, but even if it should produce promising results, pursuance of the study must be evaluated in the light of immediate benefits to present Russian-English MT.

By comparison with the preceding, the construction NOUN + NOUN IN THE DATIVE is quite straightforward. The vast majority of these constructions is derived from verbal units. Government examples are furnished by the nominalizations POMO56 DRUGU (AID TO A FRIEND) from POMOGAT6 DRUGU (TO HELP A FRIEND); SOVET SESTRE (ADVICE TO SISTER) from SOVETOVAT6 SESTRE (TO ADVISE SISTER); POTVORSTVO REBENKU (INDULGENCE TOWARD A CHILD) from POTVORSTVOVAT6 REBENKU (TO SHOW INDULGENCE



NOTE: Subscript n = nominative
Subscript g = genitive
Superscript c = concrete
Superscript $\bar{c}$ = non-concrete

Figure 2.2-5 Nominal Constructions with the Genitive.

TOWARD A CHILD). Here too there are nominalizations directly derived from adjectival units but indirectly related to verbal units such as VERNOST6 IDEALAM (FAITHFULNESS TO IDEALS) directly from VERNY1 IDEALAM (FAITHFUL TO IDEALS) and indirectly from VERIT6 IDEALAM (TO HAVE FAITH IN IDEALS). Among the inherent types there are certain interesting constructions where the original verbal unit requires an accusative. Compare the following dative constructions and their original verb phrases; UPREK DRUGU (REPROACH TO A FRIEND) from UPREKAT6 DRUGA (TO REPROACH A FRIEND) and POXVALA BRATU (PRAISE FOR THE BROTHER) from POXVALIT6 BRATA (TO PRAISE THE BROTHER).

The nominalization NOUN + NOUN IN THE INSTRUMENTAL is somewhat more widespread than the preceding. Inherent or non-derived constructions are limited semantically and usually indicate resemblance of the first noun to the noun in the instrumental, e.g., BORODKA KLINOM (GOATEE = BEARD LIKE A WEDGE) or VOLOSY EJIKOM (CREW CUT = HAIR LIKE A HEDGE HOG).

Derived constructions with a noun in the instrumental are very numerous, for any of the multifarious instrumental usages may appear with appropriate verbal nouns. Note the following types of nominalizations and their antecedent verb phrases: KOMANDOVANIE POLKOM (COMMAND OF A REGIMENT) from KOMANDOVAT6 POLKOM (TO COMMAND A REGIMENT); BURENIE VODO1 (WATER BORING) from BURIT6 VODO1 (TO BORE WITH WATER); EZDA AVTOMOBILEM (DRIVING BY CAR) from EZDIT6 AVTOMOBILEM (TO DRIVE BY CAR).

The prepositional government of nouns is fully as varied as that of verbs. But a few examples will suffice to demonstrate their multiplicity and variety. Some examples of the so-called non-derived type are:

GENERAL IZ SOLDAT	(A GENERAL FROM THE RANKS)
BANKA IZ-POD VAREN64	(A TIN OF JAM)
SYROST6 OT ZEMLI	(DAMPNESS FROM THE GROUND)

KVARTIRA V P4T6 KOMNAT (A FIVE-ROOM APARTMENT)
SITEQ NA RUBAWKI (CLOTH FOR SHIRTS)
DEN6 PERED VYXODOM (DAY BEFORE DEPARTURE)
QERKOV6 V DEREVNE (CHURCH IN THE VILLAGE)

Again such non-derived types would not become candidates for a compilation effort because of their great variety and the seemingly random nature of the combinations.

Derived constructions, however, are prime candidates for data gathering since they closely parallel the structure of verb phrases. A few examples are appended below, together with their corresponding verb phrase:

ZABOTA O CELOVEKE (CONCERN ABOUT A PERSON)

from

ZABOTIT6S4 O CELOVEKE (TO BE CONCERNED ABOUT A PERSON)

SMEX OT RADOSTI (LAUGHTER FROM JOY)

from

SME4T6S4 OT RADOSTI (TO LAUGH FROM JOY)

VXOD V ZOOLOGICESKI1 SAD (ENTRANCE TO THE ZOO)

from

VXODIT6 V ZOOLOGICESKI1 SAD (TO ENTER THE ZOO)

Nominal government of infinitives clearly displays instances of derived constructions. Note the following examples:

POPYTKA VLEZT6 (ATTEMPT TO CRAWL IN)

from

POPYTAT6S4 VLEZT6 (TO TRY TO CRAWL IN)

JELANIE RABOTAT6 (DESIRE TO WORK)

from

JELAT6 RABOTAT6 (TO WISH TO WORK)

BO4ZN6 VYXODIT6 (FEAR OF GOING OUT)

from

BO4T6S4 VYXODIT6 (TO FEAR GOING OUT)

2.2.3.5 Adjectival Government. Although adjectival government is statistically less frequent in texts, it is to be reckoned with in a data-gathering operation as an important syntactic relationship, the recognition of which can solve some common ambiguities within Russian sentences. As indicated previously, adjectival government is significantly related to both verbal and nominal government, and gathering these data in mass should be a required first step toward their proper study. Adjectives in Russian may govern nouns in various cases, prepositional phrases, and infinitives. These three features of government will be illustrated below in a straightforward manner without the rather artificial distinctions of inherent versus derived and strong versus weak government.

Adjectival Government of Nouns

Genitive Case

The most frequent usage is exemplified by the genitive after the comparative form of the adjective signifying the object of comparison.

SIL6NEE MEN4	(STRONGER THAN I)
BYSTREE ZA1QA	(FASTER THAN A HARE)
BELEE SNEGA	(WHITER THAN SNOW)

In addition there is a series of Russian adjectives that inherently governs the genitive case. A few of these are listed below.

POLNY1 VODY	(FULL OF WATER)
DOSTO1NY1 FOXVAL	(WORTHY OF PRAISE)
CUJDY1 SOMNENI4	(FOREIGN TO DOUBT = WITHOUT DOUBTS)

Dative Case

A few adjectives fall into this category, for example:

VERNYI SLOVU (TRUE TO HIS WORD)

PODOBNYI MOLNII (SIMILAR TO LIGHTNING)

PODLEJA5II PERESMOTRU (SUBJECT TO RE-EXAMINATION)

Instrumental Case

Very few adjectives may be classified here. A few examples are the following:

IZVESTNYI SVOIMI RECAMI (KNOWN FOR HIS SPEECHES)

SIL6NYI DUXOM (STRONG IN SPIRIT)

Adjectival Government of Prepositional Phrases

Many adjectives suggest and even demand an accompanying preposition together with a noun in a given case. The examples below give just a hint of the range and nature of this type of government.

Preposition DL4:

POLEZNYI DL4 DETEI (USEFUL FOR CHILDREN)

XARAKTERNYI DL4 JEN5IN (CHARACTERISTIC OF WOMEN)

Preposition DO:

MOKRYI DO KOLEN (WET TO HIS KNEES)

VESELYI DO SAMOI STAROSTI (MERRY TO A VERY OLD AGE)

Preposition IZ:

LUCWA4 IZ BRIGAD (THE BEST OF THE BRIGADES)

Preposition OT:

USTALYI OT XOD6BY (TIRED FROM WALKING)

Preposition PO:

PROSTOI PO USTROISTV (SIMPLE IN COMPOSITION)

NE PO VOZRASTU UMNYI (CLEVER BEYOND HIS YEARS)

Preposition K:

RAVNODUWNY1 K TEATRU (INDIFFERENT TO THE THEATRE)

SPOSOBNY1 K 4ZYKAM (ADEPT AT LANGUAGES)

Adjectival Government of Infinitives

Very few instances of this type of government occur among adjectives. Only two examples appear below:

GOTOVY1 SKAZAT6 (READY TO SAY)

SKLONNY1 WUTIT6 (PRONE TO JOKE)

2.2.3.6 Impersonal Constructions. If a large-scale data-gathering operation is to be seriously considered for government information, the legitimate areas of inquiry within government proper have already been discussed. Since the concept of government ramifies deeply into the grammar of a language, it remains to determine whether there are other blocks of information that may be suggested by government and that may also yield nicely to large-scale data gathering. The first candidate for consideration is the subject relationship of the various verbs that would be examined. That is, does the verb require a subject, and if so, what information can be reasonably and efficiently gathered about the subject? The first question is certainly a very important one for verbs and even sentence structure and the answer runs the gamut from sentences, usually in artistic literature, where the subject is deliberately dropped, through imperatives, where the subject is usually omitted, through first and third personal plural forms, where the subject is omitted for the expression of certain hortatory and impersonal locutions, to bona fide impersonal sentences. While all these types of subjectless sentences must be recognized in machine translation, the only type of interest for data gathering and study is the last -- bona fide impersonal sentences. For this purpose verbs must naturally be separated into 1) those which are only personal, that is, must have a subject, 2) those which may be either personal or impersonal, and 3) those which are only impersonal. Of these three categories, number three is not of immediate interest, number one is of some interest, but number two is of great interest because of the relationships between these impersonal sentences and their personal counterparts. Category number three usually contains verbs describing natural phenomena plus a handful of others. For example,

SMERKAETS4	(IT IS GROWING DARK)
RASSVETALO	(IT WAS GROWING LIGHT)
MEN4 TOWNIT OT 3TOGO	(THIS NAUSEATES ME)

Category number two is of fundamental interest to the verb system in general, and verbs in this category should constitute an important segment of Russian grammar. At the moment there seem to be two general types of such impersonal constructions formed from reflexive verbs, on the one hand, and from non-reflexive verbs, on the other hand. Reflexive impersonal verbs appear in the following kinds of sentences:

XOROWO JIVETS4 TAM	(THERE IS GOOD LIVING THERE)
EMU NE P6ETS4 SEGODN4	(HE IS NOT DRINKING TODAY)

To these examples of reflexive verbs might be added sentences like the following:

MNE KAJETS4, CTO ON BOLEN	(IT SEEMS TO ME THAT HE IS ILL)
EMU PRIXODITS4 PRIITI VO-VREM4	(HE MUST COME ON TIME)

but it is not at all clear that such usages are truly impersonal. Such instances can be elucidated only by a comprehensive study of Russian grammar.

Non-reflexive impersonal verbs appear in sentences of the following types:

LUNU ZAKRYLO OBLAKAMI (THE MOON WAS COVERED BY CLOUDS)

RYBU UBILO XOLODOM (THE FISH WAS KILLED BY THE COLD)

EMU ZALILO PODVAL (HIS BASEMENT WAS FLOODED)

RANU EMU ZAT4NULO (HIS WOUND HEALED)

EMU UNESLO VETROM GAZETU (THE WIND CARRIED HIS PAPER AWAY)

U NEGO ZVENIT V UWAX (HE HAS A RINGING IN HIS EARS)

From the scientific grammatical point of view such impersonal sentences are related in different ways to corresponding personal sentences. Thus, the sentence

EMU UNESLO VETROM GAZETU

is easily connected with the more basic personal sentence:

VETER UNES EMU GAZETU (THE WIND CARRIED HIS PAPER AWAY)

On the other hand, the sentence RANU EMU ZAT4NULO cannot apparently be related to a personal sentence like the above because there is no explicit agent for the action. The only other related and well-formed sentence possible is the following:

EGO RANA ZAT4NULAS6 (HIS WOUND HEALED)

But in this latter sentence a reflexive form is used, which suggests that there must be some more basic form. Clearly, further study of these types of sentences and their interrelationships is indicated.

From the more practical point of view of data gathering there is a series of questions to be asked about such impersonal sentences if normal word order is assumed. Can any word at all precede the verb? If a word may precede the verb, is it a noun in the dative case or a noun in the accusative or both in any of the four cases just inquired about? Can the verb combine with an infinitive of another verb? Under the same circumstances, can the impersonal verb be used with a CTO clause?

If the verb can combine with a noun in the dative or a noun in the accusative or both, may it also acquire a noun in the instrumental? All these facts need to be gathered for subsequent study even if the information should be discarded later as superfluous.

Among verbs that require a subject or that may combine with a subject, it seems feasible to inquire about the nature of the subject in the course of data compilation. The inquiry could easily get out of hand and lead to uncontrolled speculations about the nature of the universe, but it can profitably be limited to three broad categories.

1. Verbs which can be used only with animate subjects (including cases of personification which would have to be studied), e.g., CITAT6 (TO READ), PISAT6 (TO WRITE), LGAT6 (TO LIE).
2. Verbs which have no preference in regard to animate or inanimate subjects, e.g., WUMET6 (TO MAKE NOISE), STUCAT6 (TO POUND), IDTI (TO GO).
3. Verbs which can be used solely with inanimate subjects, e.g., MOROZIT6 (TO FREEZE), SVETAT6 (TO DAWN).

In the case of inanimate subjects it may prove interesting to force the classification a bit further to include the distinction between concrete and abstract nouns even though the application of these labels is not absolutely clear. Perhaps further study will clarify the essence of this distinction and permit it to be applied with assurance. It does not seem possible at the present time to venture any more detailed classification of nominal subjects, but the possible further classification of nouns in general will be raised again in this general discussion.

2.2.3.7 Translation. A tremendously important aspect of data gathering for government relationships in Russian-English MT is the recording of the preferred English equivalents to be associated with the various patterns of government. A search strategy that can locate these patterns with efficiency and accuracy will be in a position then to contribute both grammatically and semantically in a substantial way to the improvement of the output text. Instead of having to rely on "backup" entries with generalized English meanings, the translation routine will have the power to make thousands of specific choices of English meanings facilitated solely by the recognition of grammatical clues.

The verb BROSAT6 with the general meaning of TO THROW offers a good case in point. A first approximation toward a definition of its semantic sphere and its corresponding equivalents in English and a reasonable goal in terms of present knowledge about such aspects of Russian-English translation could make the following statements about this verb:

BROSAT6	+	Accusative Object	=	throw
BROSAT6	+	Instrumental Object	=	throw
BROSAT6	+	Infinitive	=	give up

The corresponding reflexive form of this verb could be preferentially translated according to the following patterns:

BROSAT6S4	+	Instrumental Object	=	throw
BROSAT6S4	+	Accusative Local Expression	=	throw oneself

Some information has already been gained by the use of a minimum and easily specified amount of government information. If this basic information can be intelligently extended to cover areas suggested by government such as information about subjects and objects the semantic specification can become more nearly precise:

$$\begin{matrix}\text{Animate}\\ \text{Inanimate}\end{matrix} > \text{Subject} + \text{BROSAT6} + \begin{matrix}\text{Animate}\\ \text{Inanimate}\end{matrix} > \text{Object in Accusative} = \text{THROW}$$

$$\text{Animate Subject} + \text{BROSAT6} + \begin{matrix}\text{Animate}\\ \text{Inanimate}\end{matrix} > \text{Object in Accusative (remarks, invectives, etc.)} = \text{HURL}$$

Animate Subject + BROSAT6 + $\frac{\text{Animate}}{\text{Inanimate}}$ > Object in Accusative = (people, positions, possessions, etc.) ABANDON

Animate Subject + BROSAT6 + Inanimate Object in Instr. = THROW

Animate Subject + BROSAT6 + Infinitive = GIVE UP

Animate Subject + BROSAT6 + Inanimate Object in Acc. = GIVE UP

Animate Subject + BROSAT6 + Inanimate Object in Acc. (activities, verbal nouns) = GIVE UP

Animate Subject + BROSAT6 + Inanimate Object in Acc. (money, etc.) = THROW AWAY

Likewise, study of the other attributes of the reflexive verb beyond its gross features of government could help in pinpointing the following meanings:

Animate Subject + BROSAT6S4 + Inanimate Object in Instr. = THROW

Animate Subject + BROSAT6S4 + Inanimate Object in Instr. (remarks, abuses, etc.) = HURL

Animate Subject + BROSAT6S4 + $\frac{\text{Inanimate}}{\text{Animate}}$ > Object in Instr. = DISDAIN

Animate Object + BROSAT6S4 + Local Prepositional Phrase in Acc. = RUSH INTO/AT

Animate Object + BROSAT6S4 + Personal Prepositional Phrase in Acc. = THROW ONESELF AT/ON

Inanimate Object + BROSAT6S4 + (Animate Object in Instr.) = BE THROWN

Inanimate Object + BROSAT6S4 + (Animate Object in Instr.) = BE HURLED (remarks, invectives)

$\frac{\text{Animate}}{\text{Inanimate}}$ > Object + BROSAT6S4 + (Animate Object in Instr.) = BE ABANDONED

Inanimate Object + BROSAT6S4 + (Animate Object in Instr.) = BE THROWN AWAY (money, etc.)

It seems clear that some kind of auxiliary information, at least, is absolutely necessary for the continued progress and improvement of MT. This information should be based on sound grammatical and semantic theory, if at all possible. It would seem ill-advised then in a data-gathering operation to have investigators at this stage indulge in semantic and grammatical speculation beyond imposing labels like animate/inanimate and concrete/abstract. The area of specific lexical items as governed elements involves mere recording without speculation. Such lexico-syntactic entities may be termed pseudo-idiomatic sequences and are of great importance to the practical solution of source-target semantics. They must be included as data to be gathered. The verb BROSAT6 discussed above offers a series of interesting examples of such pseudo-idiomatic sequences:

BROSAT6	4KOR6	=	DROP ANCHOR
BROSAT6	ORUJIE	=	THROW DOWN ARMS
BROSAT6	TEN6	=	CAST A SHADOW
BROSAT6	VZGL4D	=	DART A GLANCE
BROSAT6	JREBI1	=	CAST LOTS
BROSAT6S4	V GLAZA	=	BE STRIKING

If a large-scale government data compilation effort is put into effect, provision must be made for the efficient recording by investigators of both of the above types of semantic information.

2.2.3.8 Nominal Classification. The above discussion of translation touched upon the improvements to be gained in output by exploiting the differences in subjects and objects of verbs. These differences were to be limited to animate/inanimate and abstract/concrete relationships. The examples quoted above from the usages of the Russian verb BROSAT6, however, indicate that finer distinctions of a semantic nature are clearly desirable. It would seem reasonable that these semantic distinctions should take the form of semantic correlations that evoke in the minds of hearers the intended meaning of polysemantic items within any given sentence. These semantic correlations will have to be discovered by a long study of sentences with such polysemantic elements, sentences in which the grammatical structure is recognized as such so that it will not interfere with the semantic study. And the semantic correlations will have to be based on a set of fundamental semantic distinctions for the various elements in the sentence. A reasonable set of semantic distinctions or labels has been elaborated for Russian nouns. The set is not definitive, but it has proven useful in a number of tests. However, these tests have not been run with semantic distinctions applied to the verbs also, and they have not been made with a profound specification of Russian grammar.

2.2.3.9 Procedure for Data Gathering. In this section there appears a specific but initial suggestion for how the multifarious and extensive information discussed above might be most efficiently gathered by competent Russian grammarians. This method not only specifies just what kind of information is to be sought, but also defines how and in what order the information is to be tested for and recorded. The succeeding paragraphs present in detail a suggested procedure for the mass compilation of verbal government data. Similar procedures for adjectival and nominal government will be worked out at a later date.

All of the information discussed above can be gathered in one operation. While the volume of information is appreciable, it appears that when broken down logically, the gathering of information should not prove to be too difficult for a native speaker of Russian.

After a thoroughgoing consideration of the problem and some actual tests, it appears that the gathering of information could best be accomplished by means of a questionnaire. The information obtained from such a questionnaire can then be punched on IBM punch cards and processed by machines. Utilization of clerical help and machine time will help achieve greater economy of actual costs and speed of analysis. Extensive use of machines will also reduce the margin of errors and will be helpful in compressing the classifications of data.

For the purposes of gathering, all requests for information have been reduced to as simple routines as possible. Since the breakdown of information is very detailed, a step-by-step explanation of the proposed grammar analysis questionnaire is presented below.

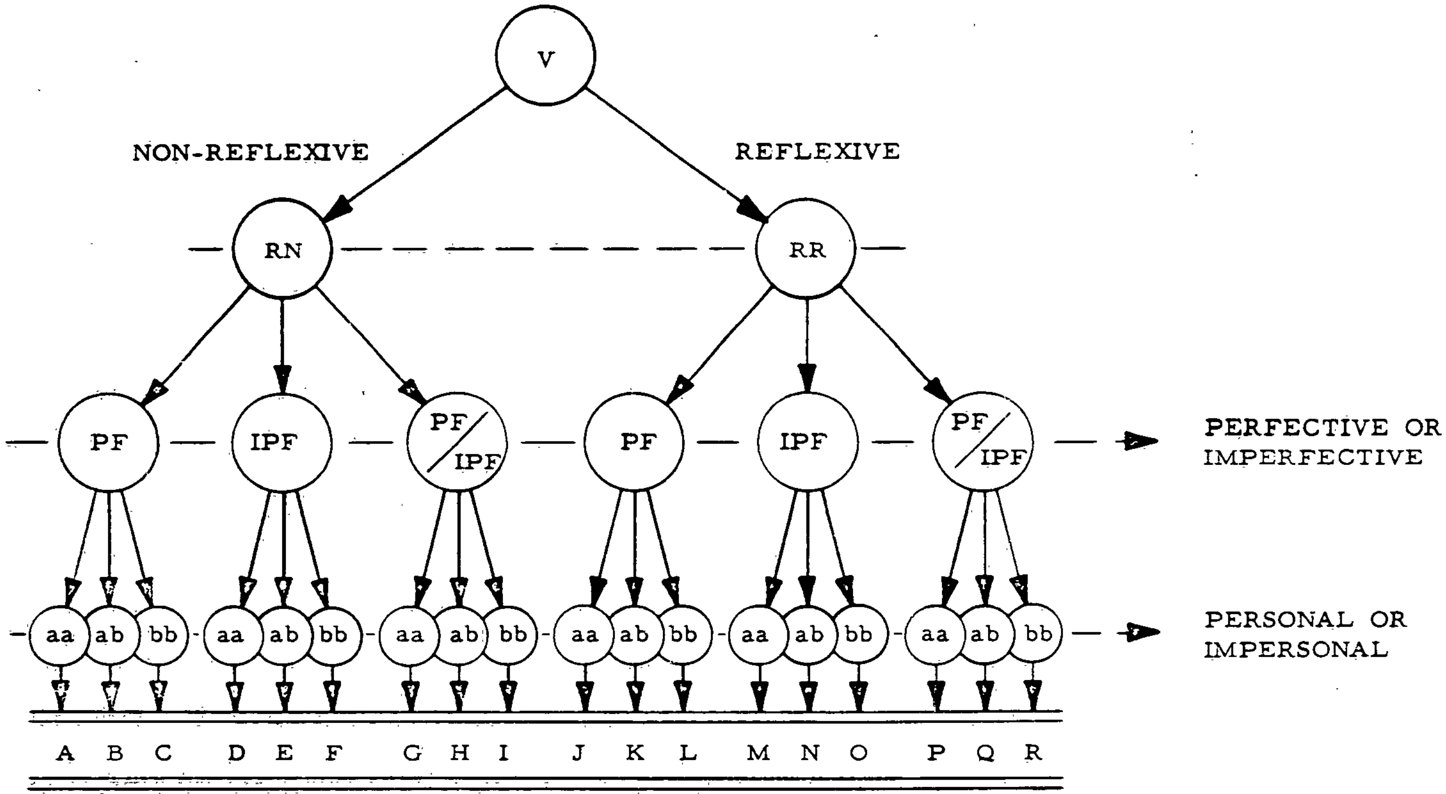
STEP I

In the upper left-hand corner of the questionnaire, write down the infinitive form of the Russian verb which is to be analyzed. Next, enter as many basic English equivalents as are necessary for adequate translation. Use lines 1 - 5. Be sure to number any additional meanings accordingly.

While it will be necessary to differentiate the various meanings as much as possible, at this point one should concentrate only on the very basic meanings, leaving idioms and phrases until the last. Thus, the verb ZARABOTAT6 will at this point be entered as <u>to earn</u> and <u>to start working</u>. Taking <u>the most general meaning</u> of the verb -- <u>to earn</u> -- one should proceed to the next step.

STEP II

Use Figure 2.2-6 to determine the classification A-R. ZARABOTAT6 in its first meaning is a non-reflexive verb, is perfective,



NOTE:
aa = verb used only impersonally
ab = verb used both personally and impersonally
bb = verb used only personally
PF = perfective
IPF = imperfective

Figure 2.2-6 Reflexivity, Aspect and Usage of Verbs.

and can be used only personally. Accordingly it shall be classified as "C" and this information recorded in the space labeled "Figure 2.2-6" in the questionnaire. The line where this information is entered should correspond to the number of the English meaning.

Note: Information contained in Figure 2.2-6 can in part be double-checked against the confix of the verb and, as suggested earlier in the report, responses against Figures 2.2-7 and 2.2-8. When the information is punched on cards, this can be done mechanically leaving for human control only those cards which are rejected as incorrect. Such checks are very important for quality control and will provide objective criteria for judging the performance of any given analyst.

The significance of Figure 2.2-6 lay in determining the very basic characteristics of the verb: reflexivity and aspect. The information regarding transitivity has purposely been omitted since the information required for analysis of government is more fully reflected in Figure 2.2-9 and the distinction between transitive-intransitive verbs can be derived automatically on the basis of information in Figure 2.2-9 and the confix of the verb.

Inclusion of the information concerning personal-impersonal usage of the verb, as mentioned earlier in the discussion, will provide valuable help in syntactic analysis. More significant is the actual economy in coding which will be possible by avoiding much of the redundancy occurring without this distinction. Finally, in the course of the analysis it will provide an important aid in differentiation of meanings.

STEP III

Depending on the responses obtained, the verb should be tested for its behavior in personal and/or impersonal usage, as the case may be. The results obtained should then be entered in spaces labeled "Figure 2.2-7" and "Figure 2.2-8" in the questionnaire. ZARABOTAT6

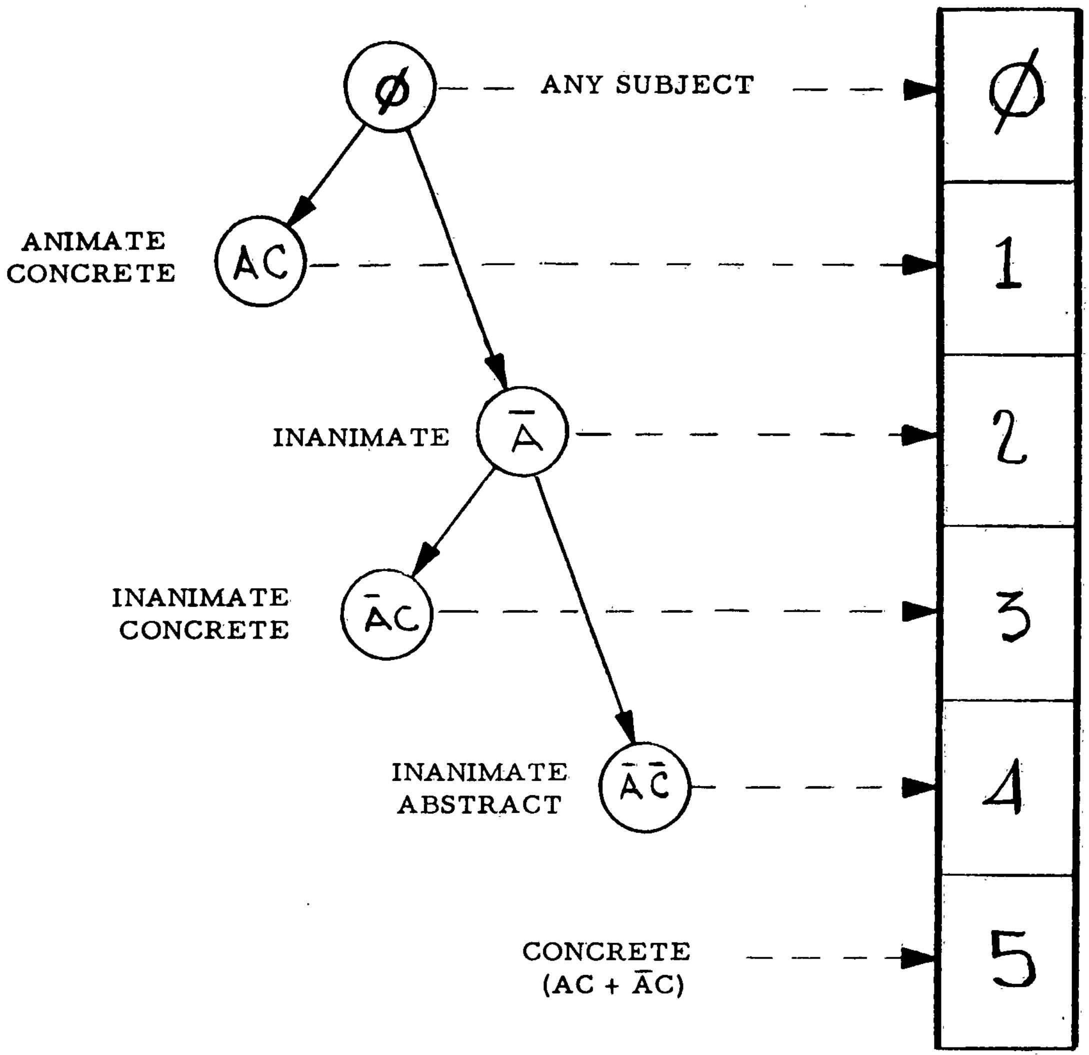
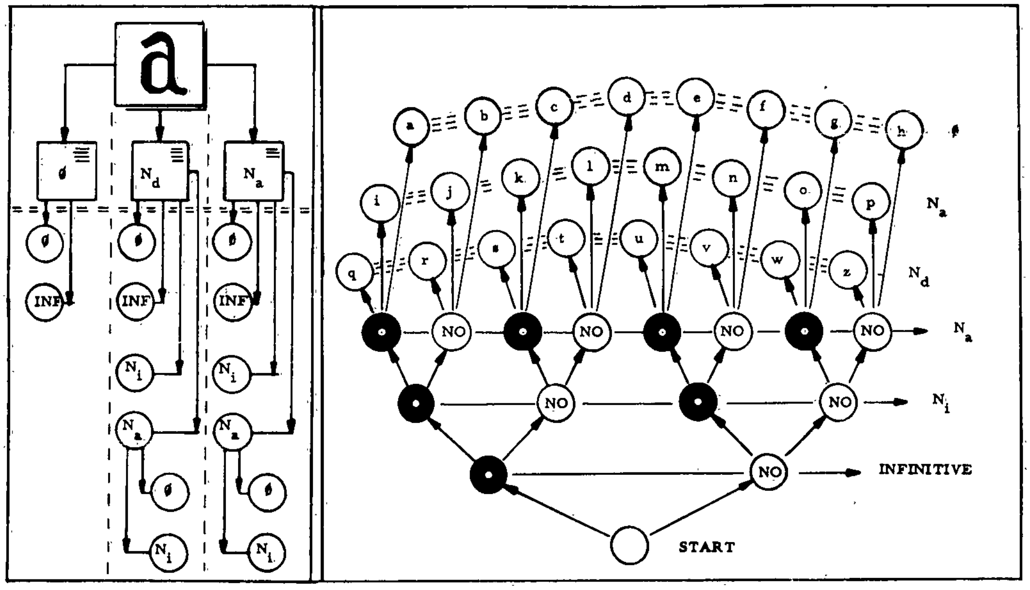


Figure 2.2-7 Subject Preference in Predication.



DIRECTIONS:

1. Take the neuter of the past tense of the verb.

2. Can the verb be preceded by a noun or a pronoun in the dative or the accusative? If it can, follow through each of the respective categories (N_d or N_a) and disregard 0.

3. Can the verb govern an infinitive of another verb? Note this fact alone and be sure not to confuse the government of the verb for that of the infinitive.

4. Can the verb, in addition, be followed by a noun in the accusative and instrumental? If yes, disregard other categories; if only the accusative is possible, mark the space marked "0" below N_a. If only the instrumental is possible, mark the space N_i above N_a in the respective column. If neither is possible, mark the space marked "0" in the respective column immediately above that which is marked "INF."

5. Trace the information obtained in the diagram on the right. Black circle is "YES," blank circle -- "NO." Follow the diagram and record the letter. Repeat the same for other column. One or two letter code will be the designation for data.

Figure 2.2-8 Impersonal Constructions of the Verb.

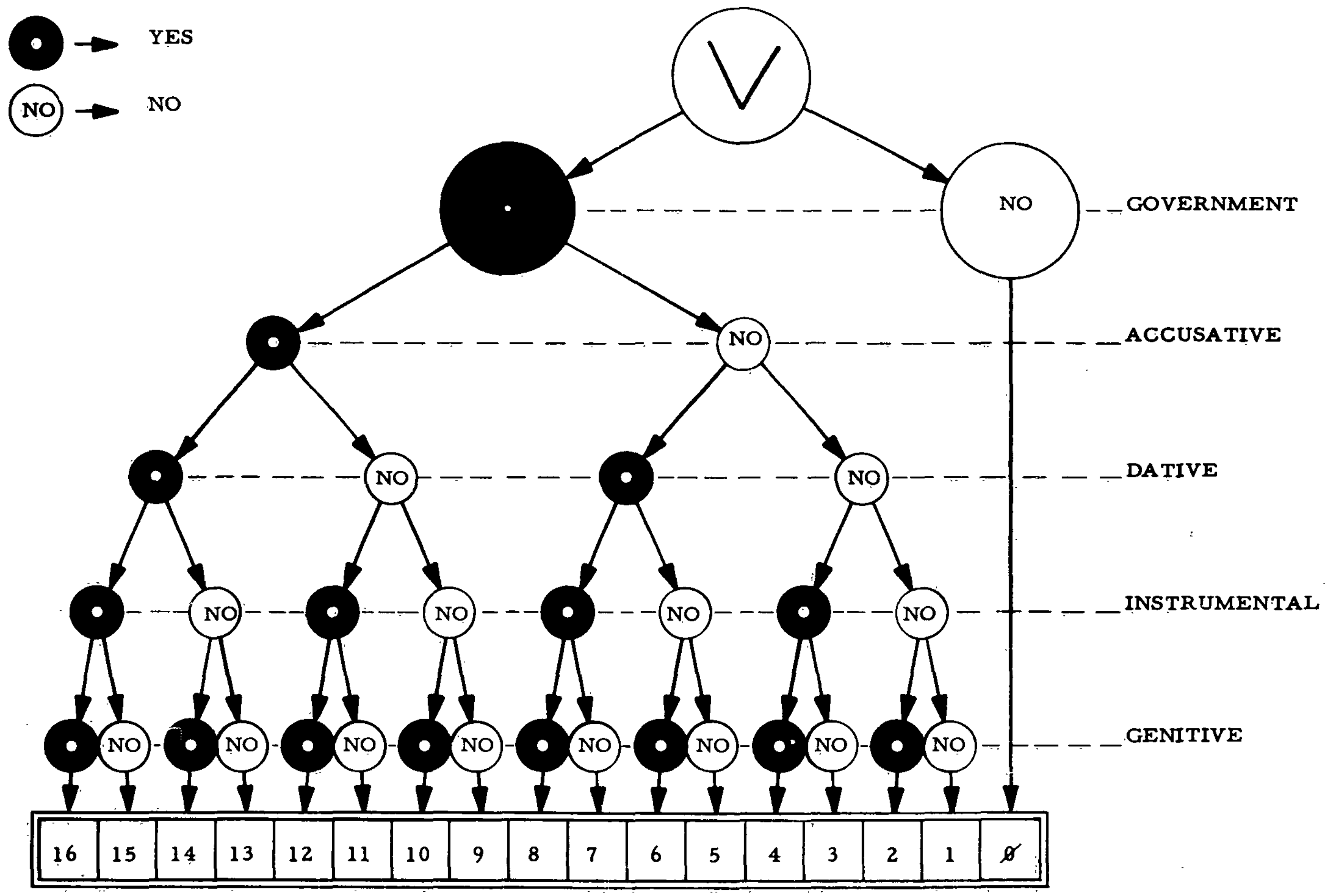


Figure 2.2-9 Direct Nominal Government Capabilities of Verbs.

can only be found in personal constructions* and there is no subject preference **. Accordingly a "$\emptyset$" shall be entered in the appropriate place. Any information request which does not apply should be left blank.

Note: In addition to the significance of Figures 2.2-7 and 2.2-8 mentioned above, the information contained in Figure 2.2-8 can be further correlated to some of the data in Figure 2.2-9, thus improving the precision of analysis.

Although we have alluded to information presented in Figures 2.2-7 and 2.2-8 earlier, Figure 2.2-8 is important for classifying and identifying impersonal usage and will help to resolve the inversion often co-occurring with impersonal usage of the verb. Moreover, it will suggest but not necessarily duplicate the information in Figure 2.2-9.

STEP IV

At this point, as described in Figure 2.2-9, one should establish the type of direct government possible with a given verb. The diagram should be read all the way down and the number will then describe the type of construction possible. Eventually the number of possible combinations will be reduced but the diagram is more manageable in its present form.

Since ZARABOTAT6 could appear in a sentence like ON ZARABOTAL SEBE DENEG/DEN6GI T4JELYM TRUDOM ("He earned for himself some money/money by hard work"), depending on how the partitive gentive is treated in the program, one will classify the

---

* While it may be argued that in a sentence like ZARABOTANO RABOCIMI (Earned by the workers) we have an impersonal construction, the sentence should be considered elliptical and a transform of RABOCIE ZARABOTALI (The workers earned).

** Although we classify ZARABOTAT6 as having no preference for a subject, it does, in fact, take animate subjects and falls into the category of verbs which may later help in determining the functioning of personification in the Russian language -- a factor which might help to improve grammatical analysis by machines.

verb in question # 16 or # 15. In contrast, the second meaning of ZARABOTAT6 would be classified in this step as # 1 since it is a form of a verb capable of government, but requiring "no object" in this particular form alone.

The significance of information presented in Figure 2.2-9 lay in outlining the effective limits of government of any one verb, thus providing one with the skeleton of possible structures which can be formed around a verb. The knowledge of the extent of government, obviously would have to be combined with other observations and rules; but, in terms of structural analysis, the data in Figure 2.2-9 present a key to the analysis of non-predicative constructions. While no effort is made to introduce strong and weak government, previous definitions still retain their validity and are implicit in further "steps" of the analysis routine.

Information obtained from Figure 2.2-9 should be entered in the questionnaire in the space so labeled. In the same breath, the analyst should determine whether a given verb can govern the infinitive of another verb. In the case of ZARABOTAT6 the answer is negative.

STEP V

The next step is to determine the translation and the type of constructions with the instrumental and the dative. The simple test proposed in Figure 2.2-10 will provide some important clues regarding the translation of such constructions and also furnish grammatical information.

The problem in the analysis of constructions with the instrumental is to determine the nature of the ties expressed in the same manner. Basically, then, one must distinguish between the instrumental expressing the tool of the action (PISAT6 KARANDAWOM -- WRITE WITH A PENCIL; RUBIT6 TOPOROM -- CHOP WITH AN AX;

...JEN5INU KUXARKO1

1. (V TO VREM4) KOGDA ONA BYLA KUXARKO1

2. KAK KUXARKU

A

3. (GOVORIL, SKAZAL), CTO ONA "KUXARKA"

4. PREDOSTAVIL DOLJNOST6 KUXARKI

...KARTY VEEROM

1. POSREDSTVOM VEERA

B

2. V FORME VEERA

...KOLESOM, STRELO1, LUPO1...

1. POSREDSTVOM KOLESA, STRELY, LUPY...

C

2. V FORME KOLESA, STRELY, LUPY...

3. KAK KOLESO, STRELA LUPA...

...DEN6GI BRATU...

1. TO BROTHER

D

2. FOR BROTHER

Note: The above information will be plotted in a tree diagram, similar to Figures 2.2-8 and 2.2-9, when more extensive information has been accumulated. Since the actual amount of combinations is very small no diagram is provided at this time.

Figure 2.2-10 Suggestive Usage of Nouns in Instrumental and Dative.

MYT6 MYLOM -- WASH WITH SOAP). The translation presents not much of a problem if the proper syntactic function of the instrumental is established. Similar to the above are the usages of instrumental in impersonal constructions which can be, however, identified more easily. Clearly recognizable is the use of instrumental in passive constructions UKAZANO DRUGOM -- POINTED OUT BY A FRIEND, SDELANO STOL4ROM -- DONE BY A CARPENTER, etc...

Adverbial uses of the instrumental are the hardest to recognize by machines. For this purpose the classification of nouns in Figure 2.2-11 will provide some helpful clues. Especially useful are nouns N_6 and N_8.

Uses of the instrumental as part of a compound predicate can be resolved on the basis of a small group of the verbs of being (about 15) which appear as link verbs in such cases,

BYT6,	SCITAT6S4,	4VL4T6S4,	STAT6
to be	to be considered	to be	to become

and others.

Cases of the use of the instrumental as an indirect object are hard to differentiate from those described above and one can best approach the task by separating a small group of such verbs as DOROJIT6 -- TREASURE, ZANIMAT6S4 -- BE OCCUPIED, BUSY ONESELF, INTERESOVAT6S4 -- SHOW INTEREST, LHBOVAT6S4 -- ADMIRE, etc. which for the most part can govern only the instrumental.

Much valuable information can be gained from the test suggested in Figure 2.2-10 and the results obtained, in conjunction with other information, will give knowledge which would make adequate translation of such constructions possible in the majority of cases.

The verb ZARABOTAT6 would be classified in Figure 2.2-10 under A-1, B-1, C-1, D-2. As suggested earlier, this information will be compressed into a single code.

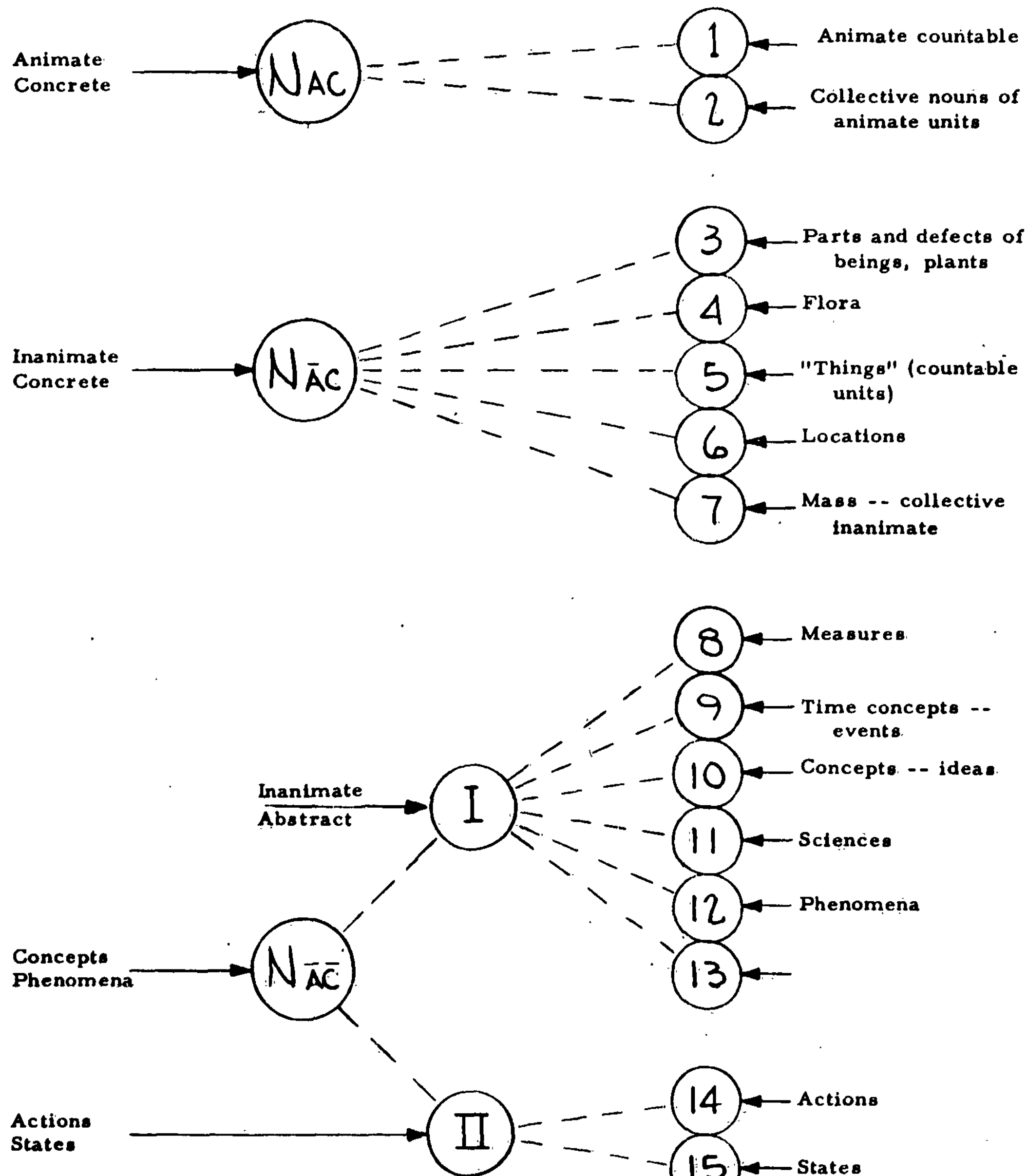


Figure 2.2-11 Tentative Semantic Classes of Russian Nouns

STEP VI

Next, the analyst should read the entries in the column on the left hand side of the questionnaire and simply mark any entry possible with the verb. The routine seems easiest if one takes any second person form of the verb and fits the various word combinations into a sentence. Numbers over columns denote English meanings.

While this aspect of the data will deal primarily with the problem of prepositional government, the information which will be derived will help to classify verbs in such a manner that not only the translation of the prepositions will be improved but much more knowledge will also be gained concerning some categories of verbs which are often mentioned in Soviet grammars but never fully enumerated. Thus # 29 will define the category of the verbs of "STRIKING." More immediately with such verbs the preposition "O" and the accusative will always be preceded and followed by a concrete noun. With the same verbs the preposition "PO" and the dative will be translated as AT when followed by a concrete noun.

Lines # 14 and 25 provide information about the group of verbs of "communication, thought, human relations." Line # 16 will identify verbs of "seizing, removal, requisition", etc. For each of the groups one may point out further related characteristics.

While errors are possible in this portion of the analysis, our basic objective is to identify at least part of the pattern emerging out of this classification. Further study will make it possible to single out such key characteristics as will be necessary for adequate analysis.

STEP VII

Upon completing this portion of the analysis the analyst is likely to notice any significant patterns. In this step the analyst should then determine whether or not the words which a given verb governs influence the meaning of the verb. In many instances it is possible to resolve ambiguities by noting which of the nouns follow the verb. The classification of nouns as suggested in Figure 2.2-11 seems to serve well in practical application. Thus the verb, REWIT6, has two basic meanings: "to decide" and "to solve." In looking at it one will discover that when REWIT6 is followed by the infinitive the translation should be "decide," when followed by any abstract noun it should be translated as "solve."

Categories of nouns in Figure 2.2-11 which help to resolve ambiguities should be entered in the space marked "N."

Since in the case of ZARABOTAT6 the ambiguity is resolved on the basis of presence or absence of an object it is not necessary to single out any one group of nouns. Space above the heavy line next to the English meaning numbers can be used for brief comments or notes.

STEP VIII

After having resolved all possible cases in which the meaning 1 occurs, the analyst should repeat all the preceding steps for other meanings of the verb, noting in each only the variations in responses.

STEP IX

After all of the meanings have been entered, the analyst should record in the space designated as "Idioms" any peculiar idiomatic or phraseological constructions as they appear in the Russian dictionary. Other comments should be entered in the space so labeled. This concludes the work of the analyst.

FURTHER PROCESSING

All of the information recorded on the Grammar Analysis Questionnaire (Figure 2.2-12) would then have to be punched on cards

Russian Verb

English Meanings
1
2
3
4
5

COMMENTS

Mark which of the following is possible with the verb. Use second person in any tense. Do not add any additional words.

5	4	3	2	1	No.	English	Russian
					1	Before	ДО РАССВЕТА
					2	Until	
					3	As far as	ДО МЕНЯ
					4	Before	
					5	Because of	ИЗ-ЗА СТОЛА
					6	From behind	
					7	For	К МИТИНГУ, СОБРАНИЮ
					8	To	
					9	By (the time of)	
					10	To	К НАМ
					11	Toward	
					12	At (during)	ЗА ОБЕДОМ
					13	After (to get)	
					14	О СЕСТРЕ, ЗИМЕ, ТОМ	
					15	At Zina's	У ЗИНЫ
					16	From Zina	
					17	In imitation of	ПОД ПУШКИНА
					18	Under	
					19	For	ПОД КАПУСТУ, ДЕТСАД
					20	Under	
					21	ПО ВОДУ, ЯГОДЫ, ХЛЕБ	
					22	For	ЗА СТОЛ
					23	Behind	
					24	At (grab)	
					25	..., ЧТО НАПИШЕТ, ОТДАСТ	
					26	On	НА ДЕРЕВО
					27	At	
					28	In the direction of ...	
					29	В СТОЛ, СТЕНУ	
					30	ВДВОЕ, ВЧЕТВЕРО	
					31	СИЛЬНО, ОЧЕНЬ	
					32	НАДВОЕ, НАЧЕТВЕРО	
					33	ВДВОЕМ, ПО ТРОЕ	

Fig. 3.2	1		2		3		4		5		INF		N
Fig. 3.2	1		2		3		4		5		INF		N
Fig. 3.2	1		2		3		4		5		INF		N
Fig. 3.2	1		2		3		4		5		INF		N
Fig. 3.2	1		2		3		4		5		INF		N
Fig. 3.2	1		2		3		4		5		INF		N
Fig. 3.2	1		2		3		4		5		INF		N

IDIOMS

Figure 2.2-12 Grammar Analysis Questionnaire and these cards checked both for errors in analysis and for errors in processing. As was suggested earlier, much of the work can be done mechanically.

2.2.4 Russian Grammatical Studies

2.2.4.1 General Discussion Among those without technical linguistic knowledge, and even among some linguists, by far the most popular conception of sentence organization is dominated by the notion that a sentence of a natural language is nothing more than a selection of several words from a dictionary which have been placed in some serial order. While in a sense this is quite true, it is also most unilluminating, for in this view it proves to be impossible to specify just what serial order it is necessary to impose on a set of words to insure that the sequence will be a sentence of the language. Serious study of such a question generates the central core of linguistic science, the study of grammar, and there is no doubt that linguistics is in pressing need of greater clarification in this area: answers to questions about sentence structure are the foundation of all other substantive inquiries into language behavior.

According to the most advanced knowledge available about the formal structure of natural languages, to understand the distinction between sentences and nonsense strings of words it is necessary to impose very special constraints upon the organization of a finite set of rewrite rules, i.e., the grammar, which express the way in which grammatical sentences are constructed.

A reasonable requirement on a grammar is that it state exactly and without recourse to intuition (1) the rules for the construction of all sentences of a language and (2) that it be capable of stating the structure of all the sentences which it can construct. Since the set of sentences is infinite, the set of rules for constructing sentences must be either infinite or recursive. The set of sentences is certainly to be completely or reasonably completely specified, yet the set of rules must not be infinite because this would be contrary to our knowledge about grammar and contrary to our concept of an interesting grammar. The set of rules must therefore be recursive. Such a set of rules for constructing an infinitely denumerable set of sentences and for describing their structure is called 'a set of sentence-generating rules' and each rule is known as a 'generative rule.'

Another requirement we might want to place on a grammar is that it be capable of deciding whether an arbitrary string is a sentence; and, if that is so, it must be capable of stating its structure. In other words it is reasonable also to expect a grammar to be a recognition device. The ability to recognize sentences presupposes the ability to define sentences and their structure. Since the latter is the goal of sentence generating rules, the discovery of rules for sentence generation must precede the discovery of rules for sentence recognition.

The format of sentence-generating rules must take into consideration the goals set for sentence generation. These goals are:

(1) that the set of sentences generated by the rules consist of all and only sentences, and

(2) that each sentence be assigned one or more structures (a string which is assigned more than one structure is said to be grammatically ambiguous).

To generate an infinite set of sentences the rules must permit recursion. And to assign a structure to the sentences it generates, a grammar must define the structure of a sentence as a list of all generative rules which were applied in producing it. An expansion of the notions of grammar and grammatical rules and their mathematical implications will be undertaken after their practical demonstration in the succeeding pages.

Intensive studies of English syntax and other areas of formal grammar have revealed amongst many other important details that there are at least three distinct levels of grammatical representation of sentence organization and correspondingly that there must be three distinct groups or types of grammatical rule. The three types of grammatical rules required to produce the set of sentences of a language are:

a) Constituent-Structure Rules

b) Transformation Rules

c) Morphophonemic Rules

These three types of rules will be discussed in the following paragraphs.

2.2.4-2 Constituent-Structure Rules The first type of rule is so constrained that it permits the reconstruction of a "tree of derivation" (see Figure 2.2-7) or phrase-structure bracketing which is imposed on the derived sentence, and the set of these so-called "constituent-structure rules" thus provides the basic branching diagram of constituents for all sentences. Thus, by virtue of the application of these rules in the derivation of sentences within a grammar, sentences acquire an underlying constituent structure similar to that which algebraic or logical expressions have, namely, a kind of nonoverlapping parenthesization. Such rules, also termed "rules of formation," have the general form: $A \rightarrow Y$ in the environment X—— Z where $\underline{A}$ is a simple syntactic element (label of a node in a tree), $\underline{Y}$ is a specified non-null string of elements, (the expansion of $\underline{A}$ in the environment X—— Z), $\underline{X}$ and $\underline{Z}$ are possibly null strings of elements to the left and to the right, and where the arrow may be interpreted as "may be rewritten as."

By way of further explanation it should be mentioned that this type of rule, as are the others to follow, is strongly motivated by the assuredly primary notion that grammars should be regarded as sentence-enumerating algorithms, among other things, of course. These algorithms should reflect the syntactic patterning of natural-language sentences in which the presence and arrangement of certain elements must be representable in the form of concatenated strings upon which is imposed a labeled bracketing. This first set of rules, constituent-structure or formation rules, is capable of generating in any one language all sentences of a maximally single or central type, usually termed "kernel" sentences. When this set of rules, most of which are optionally applicable, is put into operation, it produces a derivation of all output kernel sentences. That is, beginning with the symbol S for sentence, each successive rule of formation, if applicable, permits the derivation of a new expanded string from a previous string by the conversion of only a single symbol at a time. This succession of derivations continues until a so-called terminal string is reached, the most detailed structural representation of a kernel sentence before the application of another set of rules - the morphophonemic rules. The entire derivation can be represented on a branching diagram or tree of derivation as in Figure 2.2-14. This labeled bracketing, so clearly represented in trees of derivation, is of great significance; for not only does it formalize directly the notion of grammatical category in the nodes of the tree structure but also its higher-level elements or nodes are very important for the subsequent application of very powerful and productive processes (rules of transformation) that serve to derive (in another sense of the word) complex sentences from underlying simple or kernel sentences.

A sample set of such rules of formation and their derivation of a kernel sentence* might be the following:

---

* This sentence was taken from some research notes produced by Prof. R. B. Lees of the University of Illinois while he was working at IBM Research.

<u>Constituent-Structure Rules</u> **(See Figure 2.2-13 for the tree structure corresponding to this set of rules)**

1. Sentence → Nominal + Verb Phrase
2. Verb Phrase → Auxiliary + [Verb + Modifiers] *
3. [Verb + Modifiers] → Transitive Verb + Nominal
4. Nominal → Noun Phrase + Number
5. Noun Phrase → Article + Noun
6. Noun → Inanimate Noun
7. Number → Singular
8. Auxiliary → Tense
9. Tense → Past
10. Transitive Verb → calculate
11. Inanimate Noun → computer, logarithm
12. Article → the

<u>Sentence Derivation by the Above Set of Constituent-Structure Rules</u>

1. Nominal + Verb Phrase
2. Nom + Auxiliary + [Verb + Modifiers]
3. Nom + Aux + Transitive Verb + Nom
4. Noun Phrase + Number + Aux + V_{tr} + Noun Phrase + Number
5. Article + Noun + N^{O} + Aux + V_{tr} + Article + Noun + N^{O}
6. T + Inanimate N + N^{O} + Aux + V_{tr} + T + Inanimate N + N^{O}
7. T + N_{in} + Sg + Aux + V_{tr} + T + N_{in} + Sg
8. T + N_{in} + Sg + Tense + V_{tr} + T + N_{in} + Sg
9. T + N_{in} + Sg + Past + V_{tr} + T + N_{in} + Sg

10.-12.

the + computer + Sg + Past + calculate + the + logarithm + Sg **

* Square brackets enclose unitary structures

** Some additional constraints, possibly semantic, would be necessary to avoid the other possible sentence: The logarithm calculated the computer.

The preceding example of sentence production was kept deliberately simple. The rules of formation were such that a symbol could be re-written only as either a single symbol or as a string of symbols.

E.g., Tense ---- → Past
Noun ------ → Inanimate Noun
Noun Phrase → Article + Noun

However, to have the phrase-structure grammar generate all simple sentences of English, we would have to include among the rules of formation a new type of rule, a disjunctive rule. Examples of such rules are given below:

E.g., Tense ------- → 'Past' or 'Present'
Noun ------- → 'Animate Noun' or 'Inanimate Noun'
Verb ------- → 'Intransitive Verb' or 'Transitive Verb + Nominal'

A new set of formation rules -- similar to the one given before, but including disjunctive rules -- is listed below. Some new labels for nodes have been introduced. These labels may not withstand close inspection, but they have proved useful in this more detailed set of rules. The nature of these labels should be clear from the tree diagram in Figure 2.2-13 which follows directly. Figure 2.2-13 is a graphic representation of this more sophisticated set of constituent-structure rules. Just after Figure 2.2-13 there follows a graph Figure 2.2-14 indicating how the derivation of the sample sentence can be mapped into the tree structure of formation rules.

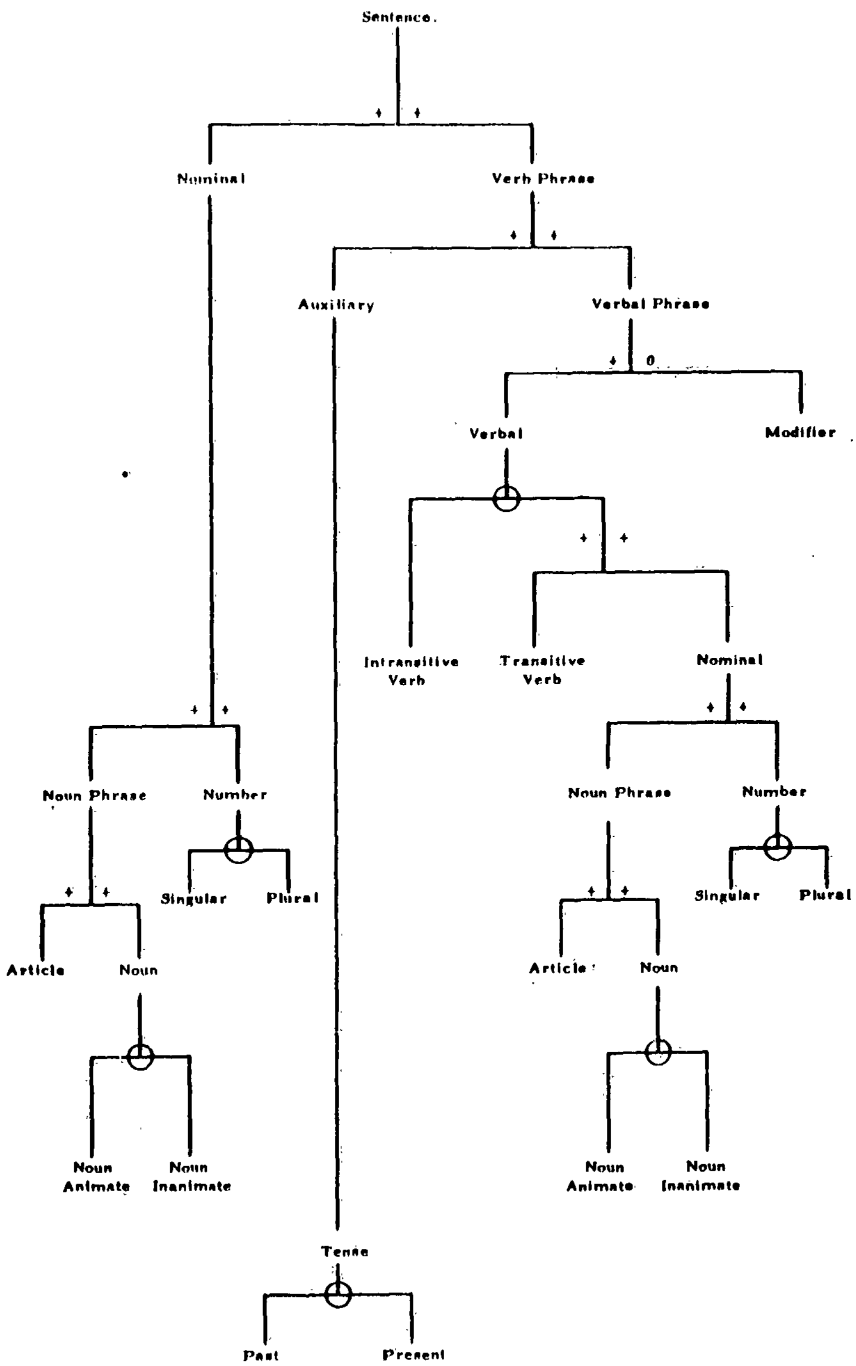


Figure 2.2-13 Graphic Representation of Constituent-Structure Rules

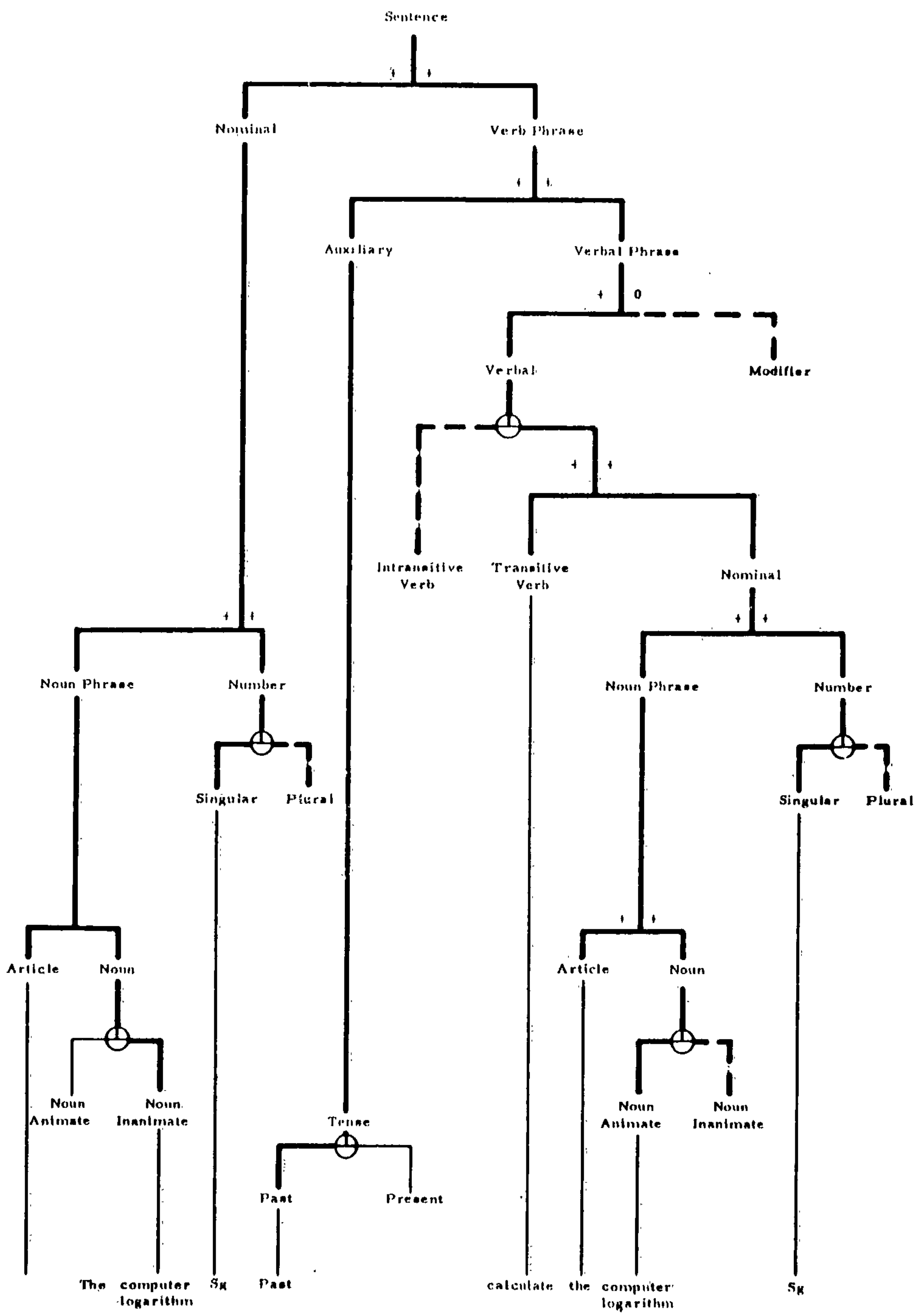


Figure 2.2-14. Sentence Derivation Mapped into the Tree Structure for Rules of Formation

Rules of Formation

(1) Sentence ------→ Nominal + Verb Phrase

(2) Verb Phrase ---→ Auxiliary + Verbal Phrase

(3) Verbal Phrase --→ 'Verbal' or 'Verbal + Modifier'

(4) Verbal---------→ 'Intransitive Verb' or 'Transitive Verb + Nominal'

(5) Nominal -------→ Noun Phrase + Number

(6) Number--------→ 'Singular' or 'Plural'

(7) Noun Phrase ---→ Article + Noun

(8) Noun ----------→ 'Animate Noun' or 'Inanimate Noun'

(9) Auxiliary ------→ 'Tense'

(10) Tense ---------→ 'Past' or 'Present'

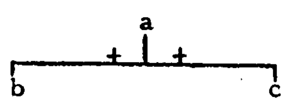
(11) Article --------→ the

(12) Inanimate Noun -→ computer, logarithm

(13) Verb transitive -→ calculate

The following explanation of symbols pertains to Figures 2.2-13 and 2.2-14.

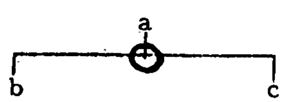
(1)



means that a is rewritten as both b and c

(a → b + c)

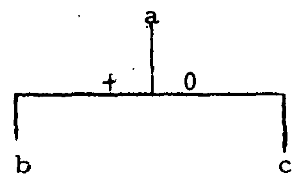
(2)



means that a is rewritten as either b or else as c

(a → 'b' or 'c')

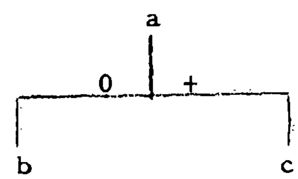
(3)



means that a is rewritten as either b or else as both b and c

(a → 'b' or 'b'+ c')

(4)



means that a is rewritten as either c or as both b and c

(a → 'c' or 'b + c')

(5) The distance of a node from the top of the diagram in Figures 2.2-13 and 2.2-14 represents the order of the rules. For example:
(1) S → Nominal + Verb Phrase is the highest ordered rule and this is shown by its being at the top of the diagram; whereas, (10) Tense →'Past' or 'Present' is the lowest ordered rule, and this is shown by its being at the bottom of the diagram.

2.2.4.3 Transformation Rules. The second type of rule is represented by a set of more complex rules which serve to convert certain underlying branching diagrams into new, less central, derived trees, such as those corresponding to complex sentences, within which are imbedded other simpler, already derived sentences. The exact specification of constraints which must be imposed on this type of rule, called the "grammatical transformation," is an important desideratum of contemporary linguistic research. These rules of transformation are particularly powerful, and their postulation seems indispensable to an explanatory theory of grammar; but their exact nature remains to be tested in intensive and extensive inquiries into grammars of a number of natural languages, and much controversy has arisen about them in linguistic circles. Transformational rules have a completely different form from the rules of formation discussed above. A grammatical transformation may be specified by an ordered triplet, D, Z, t, where D is a certain derivation tree or constituent structure, Z is a string of subjacent elements and t is an elementary transformation of the constituents of that string. Transformations may effect permutation, ellipsis, addition, etc., of elements. Furthermore, such rules serve to map trees into new trees, yielding strings with derived constituent structure. The recursive power of the grammar inheres in its transformational component, for such rules are permitted to reapply in specified order to already derived transforms.

The action of a specific transformational rule may be illustrated on the kernel sentence introduced above: "The computer calculated the logarithm." It has been demonstrated that passive sentences may be produced from underlying active sentences such as the one above through the agency of the so-called passive transformation. The passive transformation can be applied to any string that is analyzable into the following constituent structure:

Noun Phrase$_1$ + Auxiliary (Tense) + Verb Base + Noun Phrase$_2$ and will necessarily include the elements in the terminal string representing the sentence above:

[the + computer + Sg]$_1$ + Past + calculate + [the + logarithm + Sg]$_2$.

Since bracketing is equivalent to tree structure, these representations correspond to elements D, Z of the ordered triplet mentioned above.

The third element "t" will operate on each item in the constituent structure in turn to produce the following effect: 13. Noun Phrase$_2$ + Past + be + en + Verb + by + Noun Phrase$_1$ and the following is obtained by substitution of the appropriate English morphemes:
the logarithm + Sg + Past + be + en + calculate + by + the + computer + Sg

2.2.4-4 <u>Morphophonemic Rules</u> The third or bottom level of grammatical structure consists of morphophonemic, phonemic, and phonetic rules which map representations of sentences as strings of parenthesized morphemes into proper morphemic and phonological form, finally creating the most detailed kind of phonetic transcription. These rules will not be illustrated in their entirety here because they are not of immediate importance to lexical processing efforts such as automatic translation, but the morphophonemic rules will be demonstrated on the kernel sentence above and on its passive transform. The resulting strings will thus be left in the conventional orthographic shape. First, a listing of these obligatory rules:

<u>Morphophonemic Rules</u>

14. $\left\{ \begin{matrix} Sg \\ Pl \end{matrix} \right\} \rightarrow \left\{ \begin{matrix} \emptyset \\ S \end{matrix} \right\}$

15. X + affix + verb + Y → X + verb + affix + Y

16. Past → d; be + Past → was; en → d

17. $X \left[\begin{matrix} N^o \\ af \end{matrix} \right] Y \rightarrow X \oplus \left[\begin{matrix} N^o \\ af \end{matrix} \right] Y$ *

18. $X + Y \rightarrow X \# Y^{**}$

---

* The symbol + designates an affix linkage.

** The symbol # indicates a word boundary.

The kernel sentence will be transformed by the above obligatory rules in the following way:

the + computer + Sg + Past + calculate + the + logarithm + Sg

14. the + computer + $\emptyset$ + Past + calculate + the + logarithm + $\emptyset$

15. the + computer + $\emptyset$ + calculate + Past + the + logarithm + $\emptyset$

16. the + computer + $\emptyset$ + calculate + d + the + logarithm + $\emptyset$

17. the + computer $\oplus$ $\emptyset$ calculate $\oplus$ d + the + logarithm $\oplus$ $\emptyset$

18. the # computer $\oplus$ $\emptyset$ # calculate $\oplus$ d # the # logarithm $\oplus$ $\emptyset$

yielding the sentence below in the conventional orthography:

<u>The computer calculated the logarithm</u>

The passive transform of this sentence would be developed by the same rules, as follows:

the + logarithm + Sg + Past + be + en + calculate + by + the + computer + Sg

14. the + logarithm + $\emptyset$ + Past + be + en + calculate + by + the + computer + $\emptyset$

15. the + logarithm + $\emptyset$ + be + Past + calculate + en + by + the + computer + $\emptyset$

16. the + logarithm + $\emptyset$ + was + calculate + d + by + the + computer + $\emptyset$

17. the + logarithm ⊕ ∅ + was + calculated ⊕ d + by + the + computer ⊕ ∅

18. the # logarithm ⊕ ∅ # was # calculated ⊕ d # by # the # computer ⊕ ∅

yielding the sentence below in conventional orthography:

The logarithm was calculated by the computer

2.2.4-5 Theoretical Framework. After the above practical illustration of grammatical organization in terms of an English example, it seems reasonable now to launch into a brief consideration of underlying theory together with abstract representations of the notions of string, language, generative grammar, phrase structure grammar, grammatical rules, structural descriptions, etc.

Given a finite vocabulary V_T, a string is a finite-length concatenation of elements of V_T. The string is said to be over V_T. There are as many strings over V_T as integer numbers, the set of all strings has the power of enumerable infinity. This latter set is a semigroup with respect to concatenation, namely the concatenation of two strings is a string; and concatenation is associative: if the following s_i denotes strings and the + sign is the concatenation sign, then

$$(s_1 + s_2) + s_3 = s_1 + (s_2 + s_3)$$

This associativity feature allows us to write $s_1 + s_2 + s_3$ without ambiguity, omitting the parentheses.

A language L is a subset of the set of all strings over V_T. A string of L is a sentence. Interesting languages are not finite:

they have an enumerable infinity of sentences, at most. The number of languages over V_T is continuous infinity.

The generation of a language L is accomplished with the help of a finite language, called generative grammar, over a vocabulary V, such that

$$V = V_T \cup V_N, \quad V_T \cap V_N = \phi, \; V_N \neq \phi.$$

V_T is the terminal vocabulary: the vocabulary of the language L to be generated; V_N is the non-terminal vocabulary.

A sentence of the grammar is called a rule. It is essential that the number of rules should be finite, otherwise we would not gain anything by substituting the grammar for the language, as a generating device.

In order to keep a generative grammar compatible with a Turing machine, the general form of a rule is

$$s_1 \rightarrow s_2$$

where s_1 and s_2 are strings over V; and $\rightarrow$ interpreted as "rewrite as," is an element of V_N, excluded as a permissible component from s_1.

A set of rules of the above form - called an "unrestricted rewriting system" - can hardly qualify to be a grammar. In order to make this system a "rewriting grammar" we have to impose two further conditions:

1) if l (s) is the "length of s", i.e. the number of elements of V concatenated, we have to have: $l(s_1) \leq l(s_2)$. The rules have to be "length-preserving." This condition is needed for decidability purposes. In particular, deletions are forbidden.

2) s_1 must not be terminal: at least one element in s_1 must be non-terminal. This condition is needed to terminate the generative process.

A rewriting grammar is called a phrase-structure grammar if its rules are of the form

$$\phi+A+\psi \rightarrow \phi+a+\psi$$

ϕ, ψ, a are strings, and A is a non-terminal string of length one. ϕ and ψ may be null strings.

A phrase-structure grammar is context-sensitive, if in at least one of its rules ϕ or ψ or both are non-null. Otherwise, it is context-free. Thence, the rules of a context-free grammar are of the form

$$A \rightarrow a.$$

A context-free grammar is regular if all its rules are of the form

$$A \rightarrow a + B$$

where a is a single terminal element, B a single non-terminal element. Languages have the same class-names, i.e., context-free, context-sensitive, as their generative grammars. The generative process takes place as follows:

Given a grammar G and the language L to be generated, we define an intermediate language L_1 between G and L. L_1 has an infinite number of sentences, defined recursively:

1) S is a special element of V_N; the string of length one S is a sentence of L_1.
2) If $\phi+A+\psi$ is a sentence of L_1 and if $A \rightarrow a$ is a rule of G, than $\phi+a+\psi$ is a sentence of L_1.

L is the proper subset of L_1 whose sentences are all terminal. Given a sentence of L, its S-derivation is a finite sequence of sentences of L_1, the first sentence of the sequence being S, the last one of the sentence of L, and any member of the sequence is obtained from the previous one by a rewriting rule, except S, of course. Sentences of L that have no S-derivation cannot be generated. In this case the grammar is really inadequate for L.

Two grammars are weakly equivalent if they generate the same language. It is obvious from the above described generative process that a grammar generates only one language. The number of rewriting grammars is enumerable infinity. The number of languages being continuous infinity, it is obvious that most languages have no generative grammar at all.

For phrase-structure grammars, one can define the "structural description" of a generated sentence. It is a string on V whose leftmost character is S. If A_k is the k^{th} element of the string (supposedly a non-terminal character), and if $A_k \rightarrow a_k^1 + \ldots. + a_k^n$, the (k+1)-th element of the string is a_k^1. If a_k^1 is terminal, the (k+2)-nd element of the string is a_k^2. If it is not, we repeat the same procedure by developing a_k^1 in the place of A_k, - until we encounter a terminal element. As an example, consider the following S-derivation of the sentence "The small boy plays baseball today."

S	$\rightarrow$	NP + VP
NP	$\rightarrow$	Art + Adj + Noun
VP	$\rightarrow$	V + NP + Time
Art	$\rightarrow$	the, Adj $\rightarrow$ small, Noun $\rightarrow$ boy
V	$\rightarrow$	plays
NP	$\rightarrow$	Noun
Noun	$\rightarrow$	baseball
Time	$\rightarrow$	today

The notations are intuitively obvious.

The structure description is:

S+NP+Art+the+Adj+small+Noun+boy+VP+V+plays+NP+Noun+baseball+

Time+today. The structural description is clearly an embedded parenthesis system in Polish notation. Hence it can be geometrically represented by a tree. Notice that eventual context-sensitivity is not visible, unless one introduces some special notation for it; that the order in which the rules were applied is not visible either, while it is in the S-derivation; and that the grammar has to have a phrase-structure.

We can require a generative grammar to generate structural descriptions instead of sentences; the sentence can be obtained by deleting - in a final extra-grammatical rewriting procedure - the non-terminal elements in the string. Two such grammars are strongly equivalent if they generate the same set of structural descriptions. In a grammar, a sentence may have more than one structural description. It is then ambiguous.

With generative grammars we associate various kinds of linguistic automata. A generator has the grammar as input, its output is the sentences of L, if the generator is "weak", or the structural descriptions of the sentences of L, if the generator is "strong." Such a generator is of little practical value, but current theory of human and artificial intelligence asserts we have to be able to construct it, at least theoretically, before proceeding to more sophisticated linguistic automata, for a given class of languages.

A scanner's input is the grammar and a string over V_T; the output is a yes/no answer, according to whether the string belongs to the language as a sentence or not. A regular language's scanner is a finite automaton, a context-free language's scanner is a pushdown store. We know that a context-sensitive language's scanner goes beyond linear-bounded automata.

A recognition device works like a scanner, but besides giving a yes/no answer, its output is also the structural description. We usually require recognition devices to be finite transducers. They can be effectively constructed for context-free languages. Little is known about recognition devices of context-sensitive languages. Intuition and a sort of reasoning by analogy suggests the examination of Newell-Shaw-Simon list structures.

In mechanical translation we also need production devices, whose input is the structural description of the source language's sentence, plus some transfer grammar; its output is the target language's sentence.

Present mechanical translation techniques elaborate "MT - systems" that are simultaneously grammars, scanners, recognition and production devices and transfer procedures. Such systems have clearly no serious scientific pretensions, their only justification is effective translation, when they really do it, through suitable hardware.

In order to ameliorate the present quality of MT-output, a theoretical approach using results of mathematical linguistics and artificial intelligence theory seems necessary. A research schedule valid for both linguistics and MT would be:

1) Generative grammars covering an ever-extending part of the source and target languages.
2) Recognition procedures of the source language covering the part already generable by grammars and preceded by a search strategy covering the whole language.
3) Transfer and production procedures.
4) Heuristic procedure: given a huge number of sentences of L, a machine should be able to write the grammar, provided its class is specified.

We wish to underline that no heuristic procedure is possible at an earlier date. Let us illustrate this by an example:

Suppose a sentence is a correct multiplication of integer numbers, $5 \times 12 = 60$, for instance. Suppose we give a machine a huge number of such multiplications and we require it to print out the "multiplication grammar", i.e. the multiplication rules. Perhaps we could write such a program, but only because we already know the multiplication rules. Otherwise we would not even know how to make the most elementary statements. This example shows the fallacy of so-called "discovery procedures."

Present MT-systems that are really translating use the only existing grammars: traditional ones, combined with the use of sophisticated hardware and software. Present thought is, however, that context-free grammars are inadequate for the generation of natural languages. Authors having heavily contributed to the theory of context-free languages (Chomsky, Schützenberger, Bar-Hillel) warn constantly against their use in MT.

To furnish a more adequate tool for the generation of natural languages, Chomsky proposes transformational grammars. A transformation rule is of the form

$$(s_1, s_2, \ldots . s_n) \rightarrow s$$

where the s_i are structural descriptions furnished by a phrase-structure grammar. In other words, a transformational rule operates on a string sequence. Notice that the s_i on the left side are not concatenated; they are the components of a sort of a string vector.

The main advantages of transformational rules are:

1) Total satisfaction of the linguist's intuition. In fact, linguists always used transformations, until the advent of the

sterile "structuralist' school of the last 20 years. In this sense, transformational grammars are a formalized continuation of Ferdinand de Saussure's work. The research schedule that gives theoretical priority to generative grammars, before going into pragmatics (recognition and production), semantics and heuristic procedures, is also implied - as Chomsky points out, in de Saussure's "language vs language" theory.

2) Economy. Context-free grammars fail presumably because they introduce hundreds of thousands of rules and word classes in natural languages. Nobody can compile such a grammar or master it while its compilation goes on. Transformation rules are meant to reduce seriously the number of rules and classes. Besides, context-free grammars assign to sentences the tree-type structural description given above. Apart from simple, short sentences, this type of structural description is rather unsatisfactory. In order to obtain something more powerful we have to operate on these trees and replace them, by transformation, with a resulting new kind of structural description.

The phrase-structure part of the grammar, generating the simplest sentences with a tree-type structural description, is called kernel. Transformation rules operate on the kernel; at the end of the generative process morphophonemic rules (presumably also transformational) yield the sentences in their written or spoken form.

The above-described linguistic theory is a meta-theory. It furnishes a formal system. How this theoretical framework will be filled by the particular contents of a given natural language is up to the linguist. Consequently, at the present time, we have a multiple task:

1) further elaboration of the meta-theory
2) testing and improvement of the meta-theory by way of application of the theory to specific grammars of various natural languages.
3) Theoretical modelling and practical realization of a linguistic automaton having all the necessary requirements for language data processing, that is to say: generative, recognitional, transfer, and productional features.

2.2.4.6 Application to Russian. Throughout the past eight years at MIT and during the past three years at IBM Research and currently at IBM and the University of Illinois in conjunction with Professor R. B. Lees, a substantial set of rules of formation (constituent-structure rules) and rules of transformation have been elaborated for English among many other substantial inquiries into the nature of language.

Because of the cogency of the theoretical orientation and of the fresh insight which these studies have provided for English grammar, a natural outgrowth of such linguistic investigations was the initiation of an attempt to formalize Russian grammar in a similar fashion. Accordingly, such a program of grammatical research was instituted at IBM Research during the contract period represented by this final report. There is a variety of reasons why an independent study of Russian grammar is being vigorously pursued at IBM Research. From a theoretical standpoint an intensive investigation of syntactic structure is the only known independent way to gain a clearer understanding of language, machine translation, information retrieval, linguistic behavior, language learning, etc. The principal task of a linguistic scientist in the study of grammar is to specify in rigorous detail the formal structure of grammatical rules, as previously outlined. A grammar may be regarded as a kind of automaton lying, in mathematical power, between a finite-state Markov process and a universal Turing machine. The specification of the exact nature of this automaton demands the rigorous formulation and study of grammatical rules for a plurality of languages in the search for general validity of a powerful theory for natural languages and languages in general. If the linguist is required to express the grammatical rules he studies in a fashion rigorous enough for the programmer to encode for a computer, he will be led ipso facto to an exact mathematical specification of the power that must be built into the rules. Such a specification of structure and mathematical power is one of the main results of linguistic study, for it determines certain minimum requirements for lexical processing but also yields a strong implication of what kind of capabilities must be presumed to be built into the nervous system of a child to enable him to learn and use language. At the present level of linguistic investigation into the formal features of sentences and grammars, the complexity of detail of rules that must be constructed and the network of paths of derivation through the rules are such that the algorithmic and iterative features of a general-purpose computer are welcome aids to research and understanding.

From the standpoint of the problems of machine translation there are equally cogent reasons for the type of grammatical research proposed above. As research progresses in the automatic translation of texts in one language to semantically equivalent texts in another, it becomes more and more necessary to have at hand detailed knowledge of the syntactic organization of sentences in both source and target languages. Such crucial knowledge, it seems, can be gained only by compiling sentence-enumerating rules and by studying them in great detail in order to devise optimal methods for their manipulation within machines. More specifically, as techniques are perfected for assigning structural descriptions to individually presented sentences, in other words, for the automatic recognition of sentence structure, the principles of syntactic organization utilized must approach inevitably those built into any adequate sentence-generating grammar.

If the fact that grammatical research is essential to progress in machine translation is unconditionally accepted, then grammatical investigations of the above type would seem to offer the most productive framework within which to compile the necessary information. Not only have transformations proved to provide tremendous insight into grammars of individual languages by helping to specify the relationships among sentences, but also they seem to play an important role in a truly explanatory theory of language, a theory that purports to account for the linguistic behavior of a normal human being. In addition, recording the grammatical information generated in a strict rule form places correspondingly greater demands on the investigators in terms of efficiency, simplicity, and completeness and presents the data in a form already amenable to machine manipulation.

Another important point to be made about this kind of grammatical research is that a comparison of grammatical rules for both English and Russian cannot be avoided; as a matter of fact, it is rather to be encouraged and even recorded. A strict comparison of rules leads naturally to a comparison of derivations through the rules to form sentences in each language. In other words, here lies the beginnings of a comprehensive and formal study of the transfer function from a given sentence in Russian to its nearest grammatically and semantically equivalent sentence in English.

These deeper grammatical studies of Russian were initiated relatively late in the contract period and have not yet yielded substantial results. As indicated previously, the effort has concentrated on the elaboration of a substantial set of constituent-structure rules and a consideration, at least, of a similar set of rules of transformation. Many data are known and many data have been marshaled for intensive study, but only a preliminary study of the data has been completed. Because of the inconclusive nature of results obtained so far, it has been deemed expedient to illustrate the work by presenting the tentative set of rules of formation and transformation for a kernel sentence in Russian semantically similar to the English kernel sentence used previously to explicate some of the details of the general theory. It should be no surprise that the rules for English and Russian are very much alike. It should also be obvious that the majority of the differences will occur in the morphophonemic rules because of the highly inflected nature of the Russian language. The list of rules and the derivation of the Russian sentence follow.

Constituent-Structure Rules

1. Sentence → Nominal + Verb Phrase
2. Verb Phrase → Auxiliary + [verb + Modifiers]
3. [Verb + Modifiers] → Verb transitive perfective + Nominal
4. (Verb Transitive)* Nominal → Accusative + Inanimate Noun + Number
5. Nominal → Nominative + Inanimate Noun + Number
6. Number → Singular
7. Auxiliary → Tense
8. Tense → Past
9. Verb transitive perfective → VYCISLI (calculate)
10. Inanimate Noun → MAWIN (machine) + Feminine
    Inanimate Noun → LOGARIFM (logarithm) + Masculine

---

* Parentheses here indicate a contextual restriction.

The set of derivations in accordance with the application of the above rules reads as follows:

1. Nominal + Verb Phrase
2. Nom. + Auxiliary + $\ulcorner$Verb + Modifiers$\lrcorner$
3. Nom. + Aux. + Perfective Transitive Verb + Nominal
4. Nom. + Aux. + $V_{tr\ perf}$ + Accusative + Inanimate Noun + Number
5. Nominative + N_{in} + N^{o} + Aux + $V_{tr\ perf}$ + Acc. + N_{in} + N^{o}
6. Nomin. + N_{in} + Sg + Tense + $V_{tr\ perf}$ + Acc. + N_{in} + Sg
7. Nomin. + N_{in} + Sg + Past + $V_{tr\ perf}$ + Acc. + N_{in} + Sg

8-9. Nomin. + MAWIN + Fem + Sg + Past + VYCISLI + Acc. + LOGARIFM + Masc + Sg

This set of derivations can be represented by a tree structure as shown on Figure 2.2-15.

The above rules of formation have produced a string which will eventually produce a Russian sentence equivalent to "the machine calculated the logarithm." But before this sentence can be completely derived by the application of obligatory rules of transformation for morphophonemic assignments, it will be interesting to try to illustrate an optional rule of transformation. Again it will be instructive to set up a provisional passive transformation of the above sentence. The tentative general rule might read as follows:

10. $\ulcorner$Nom. + N^{1}_{in} + Sg$\lrcorner$ + $\ulcorner$Past + $V_{tr\ perf}$$\lrcorner$ + $\ulcorner$Acc + N^{2}_{in} + Sg$\lrcorner$

$\rightarrow$ $\ulcorner$Nom. + N^{2}_{in} + Sg$\lrcorner$ + { [$\ulcorner$Past + $V_{cop\ perf}$$\lrcorner$ / $\ulcorner$Pres + $Vcop_{imperf}$$\lrcorner$] + Past Passive Participle + $V_{tr\ perf}$] + $\ulcorner$Instr + N^{1}_{in} + Sg$\lrcorner$

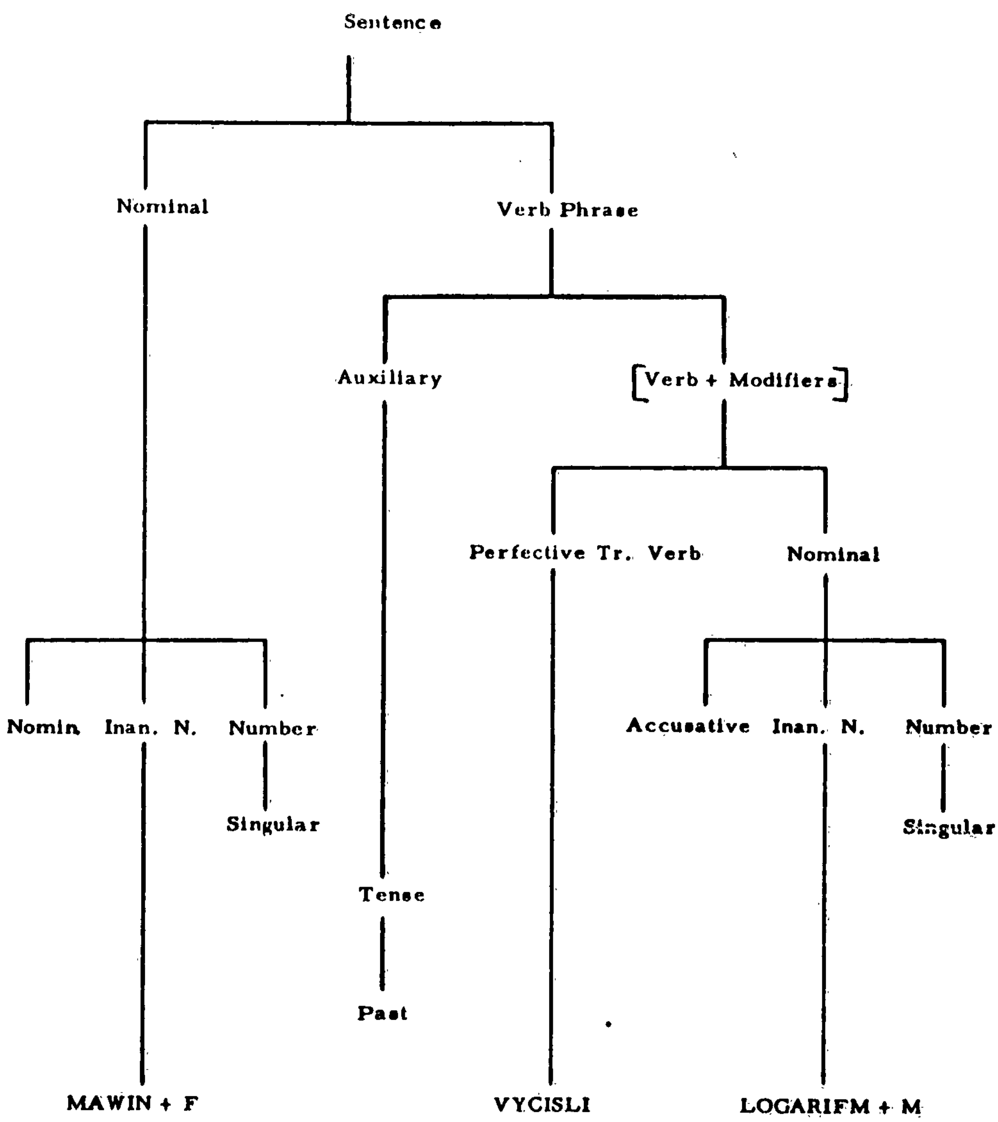


Figure 2.2-15 Derivation of the Sample Kernel Sentence

This rather complex passive transformation rule seems necessary because the passive transform of a perfective verb in the past tense requires a past passive participle preceded by the copulative verb either in the past perfective or in the present imperfective, a choice made by the speaker to fit the extra-linguistic circumstances. In the case of the sentence under discussion the past perfective form of the copula will be chosen so that the basic shape of the transform will be:

$$[\text{Nom.} + N_{in} + \text{Sg}] + [\text{Past} + \text{Vcop}_{perf} + \text{PPP} + V_{tr\ perf}] \quad +$$

$$[\text{Instr} + N_{in} + \text{Sg}]$$

Substitution of Russian morphemes for the appropriate symbols in the above string leaves the following transformed string just prior to application of the morphophonemic rules:

Nom. + LOGARIFM + Masc + Sg + Past + BY + PPP + VYCISLI +

Instr + MAWIN + Fem + Sg

Both kernel sentence and its passive transform will now be subjected to the morphophonemic rules in order to develop the present strings in an English transcription corresponding to the conventional Russian orthographic system. The morphophonemic rules are listed below.

11. $$[\text{Nom.} + N_{fem} + \text{Sg}] + \left\{\begin{matrix}\text{Past}\\ \text{PPP}\\ \text{Past} + \text{PPP}\end{matrix}\right\} \rightarrow [\text{Nom.} + N_{fem} +$$

$$\text{Sg}] + \text{Fem} \left\{\begin{matrix}\text{Past,}\\ \text{PPP}\\ \text{Past} + \text{Fem} + \text{PPP}\end{matrix}\right\}$$

$$[\text{Nom.} + N_{masc} + \text{Sg}] + \left\{\begin{matrix}\text{Past}\\ \text{PPP}\\ \text{Past} + \text{PPP}\end{matrix}\right\} \rightarrow$$

$$\boxed{\text{Nom.} + N_{masc} + \text{Sg}} + \text{Masc} + \left\{\begin{array}{l}\text{Past}\\ \text{PPP}\\ \text{Past + Masc + PPP}\end{array}\right\}$$

12. X af vY → Xv af Y

13. MAWIN + Fem + Sg + Nom. → -a

LOGARIFM + Masc + Sg + Acc → -∅

LOGARIFM + Masc + Sg + Nomin → - ∅

VYCISLI + PPP → - EN -

Past → - L -

Fem → - A -

Masc → - ∅

14. $X \left\{\begin{array}{c} N^{o} \\ af \end{array}\right\} Y \rightarrow X \oplus \left\{\begin{array}{c} N^{o} \\ af \end{array}\right\} Y$

15. $X \oplus Y \rightarrow X \# Y$

The kernel sentence would be obtained from the string in the following way:

Nom. + MAWIN + Fem + Sg + Past + VYCISLI + Acc + LOGARIFM + Masc + Sg

which was obtained by the previously stated rules 1-9:

11. Nom. + MAWIN + Fem + Sg + Fem + Past + VYCISLI+ Acc + LOGARIFM + Masc + Sg

12. MAWIN + Fem + Sg + Nom. + VYCISLI + Past + Fem + LOGARIFM + Masc + Sg + Acc

13. MAWIN + A + VYCISLI + L + A + LOGARIFM + ∅

14. MAWIN ⊕ A + VYCISLI ⊕ L ⊕ A + LOGARIFM ⊕ ∅

15. MAWIN ⊕ A + # + VYCISLI ⊕ L ⊕ A # LOGARIFM ⊕ ∅

yielding the Russian sentence below in a special transcription:

<u>MAWINA VYCISLILA LOGARIFM</u>

The passive transform of the above kernel sentence would be similarly derived:

Nom. + LOGARIFM + Masc + Sg + Past + BY + PPP + VYCISLI + Instr + MAWIN + Fem + Sg

11. Nom. + LOGARIFM + Masc + Sg + Masc + Past + BY + Masc + PPP + VYCISLI + Instr + MAWIN + Fem + Sg

12. LOGARIFM + Masc + Sg + Nom. + BY + Past + Masc + VYCISLI + PPP + Masc + MAWIN + Fem + Sg + Instr

13. LOGARIFM + ∅ + BY + L + ∅ + VYCISLI + EN + ∅ + MAWIN + 01

14. LOGARIFM ⊕ ∅ + BY ⊕ L + ∅ + VYCISLI + EN + ∅ + MAWIN + 01

15. LOGARIFM⊕∅#BY⊕ L⊕∅#VYCISLI⊕EN⊕∅#MAWIN⊕01

yielding the Russian sentence below in a special transcription:

LOGARIFM BYL VYCISLEN MAWIN01

The above sets of rules and derivations should in no way be regarded as definitive. They are based only on the data considered up to the present time. As formerly stated, they have been presented only to demonstrate the nature of the investigation to which the Russian language is being subjected and to give some idea of the eventual goals of such work. The investigation would naturally be extended throughout available Russian grammatical data and beyond so that the rules might reflect as broad a data base as possible.

2.3 System Organization

At the beginning of 1962, the AN/GSQ-16 Mark II Language Processor possessed the same system configuration that it does today. The organization of this multipass processor was thoroughly described in the final reports for the Word Analyzer, AF 30(602)-2072 and the Mark II System AF 30(602)-2080. Although no changes have been made to this system during the contract period, significant improvements have been developed and logically defined as a result of detailed studies performed during this period. Investigations have concentrated on the employment of a rapid subtable search in core memory (high-speed search), and of addenda and errata tables to supplement the main lexical tables in the Photostore.

The significance of the rapid subtable search lies mainly in its ability to perform extremely rapid processing where repeated access to limited tables is the required operation. The grammatic structure of language statements is determined in the Mark II by a contextual analysis of words employing, extensively, the type operation amenable to rapid subtable processing. In detailed estimates of production translation rates which would result from use of this procedure, it is apparent that an increase in productivity of one to two orders of magnitude may be achieved. These estimates are included in this section.

With availability of such a search algorithm as the rapid subtable search, the utilization of addenda, errata, high-frequency occurrence subtables (common words, etc.), and lexicon indices becomes both feasible and advantageous. Considerable thought has been given to the procedures required to perform language processing with such tables augmenting the Photostore, and to maintain compatibility with the longest match and integral address search algorithm. These procedures must insure that the Photostore tables are logically an extension of the subtables in core memory. Such search procedures have been largely developed so that exploitation of the addenda, errata, common words, and lexicon indices may commence as soon as system implementation is accomplished.

2.3.1 Rapid Subtable Search

Since all entries in the subtables to be processed in core memory are integrally addressed, the subtable search must be ordered in a manner similar to that employed in the Photostore so as to guarantee accession of the longest matching entry. In the Photostore a Track Search locates the point where the shortest subsequent Entry Search (serial descending entry-by-entry examination) will locate the longest matching entry existing in the lexicon. The latter Entry Search must be performed nearly identically in core memory subtables; however, the initial subtable search which is functionally equivalent to the Track Search in the Photostore may be performed by several methods in core memory. Three such methods of track or "neighborhood" searching in core memory subtables have been investigated and developed in order to ascertain an optimum routine. These are:

a Block Sampling search,

a Binary Sampling (Fibonacci) search, and

a Key Transform Directory.

The Block Sampling search is designed to effect a sampling search of the subtable in core memory in a manner almost identical to the Track Sampling search in the Photostore. Thus, most of the circuits employed for the latter would be employed for this new function. The equivalent of a track in core memory would then be composed of an arbitrary length block of entries in the subtable. Optimum access considerations have been employed to fix this block length at $\sqrt{N/6}$, where N is the total number of entries in the subtable. For this method of subtable search, the random access time, expressed in microseconds, is given by:

$$\text{access time} = 60\sqrt{N/6} + 36\ N_x$$

where N_x is the absolute minimum number of entries that must be sampled during a particular Entry search. As may be seen, access time increases nonlinearly with table size, N. For this reason, a comparison with other search methods will be contingent upon this factor.

The Fibonacci search is a modified binary sampling search wherein specific midpoint entries are selected, for a given subtable, as nodes for all possible searches. These nodal entries are provided with address vectors which specify the branch location of the next higher- and lower-valued nodal entries in the binary tree structure. For a subtable with N entries, the number of samples to be performed would be]

$$[\log_2 (N + 1)]_{\text{integer}}$$

for which the access time, in microseconds, is:

$$\text{access time} = 34[\log_2 N]_{\text{integer}} + 90\ N_x + 46$$

Thus, for an N_x of 10, the two search modes would provide an equal access time of 1.3 milliseconds for a table size of 1425 entries. Smaller tables would be searched faster by Block Sampling and larger tables searched faster by the Fibonacci search. The final selection between these search modes will obviously be determined when the subtable size has been sufficiently specified.

A Key Transform Directory method of performing the equivalent of a neighborhood search on a subtable in core memory has been developed which transforms, in a 64-word directory, a key obtained from the input text into the memory address of that neighborhood in the subtable where an Entry search should commence. The key is simply the first six-bit character of the input, and the directory, in core memory, transforms each key into one of 64 possible subtable neighborhood locations, effectively dividing the table into that many subsections. This method is extremely fast in that only two memory cycles are required to access the desired neighborhood. Here again, however, table size will be a determining factor in a final selection of search mode. Since the directory can become very large if more than 64 subsections are considered, the Key Transform Directory may not be practical for very large subtables.

2.3.2 Addenda-Errata Files

With the rapid subtable search a logical reality, some thought has been given toward incorporation of several types of subtables. Addenda are frequently necessary for a few highly specialized and novel terms which are developed in languages. Errata can be used effectively to correct both programming errors and permanent photographic errors in the Photostore. In addition, it may be advantageous to include entries for words with very frequent usage. Thus only 100 entries in the rapid-search subtable could accommodate up to 50 percent of input words, thereby considerably reducing processing time. All three of these items may be sorted together without difficulty. In some instances it will be necessary, after locating an entry in this subtable, to search the Photostore for the possibility of a longer match. Statistics for this occurrence are not yet available as they are determined by the lexicon entries on the Photostore.

Finally, the inclusion of a Photostore track index in this subtable in core memory provides a very useful means of decreasing access time in Photostore searches. Each input unit to be processed will first be compared to the rapid-search subtable, and then compared, if the matching entry in this subtable so indicates, to the Photostore for a longer match. It is very convenient to perform a track index search concurrent with this subtable search since the index entries may be sorted with other subtable entries. For this purpose, the highest valued argument on each Photostore track would serve as an entry in the subtable, with the entry function providing the Photostore track number. When an Entry search is performed in the subtable, the comparator signals will be employed to indicate the lowest valued index entry which is numerically superior to the input value. With the track location known, the Track Sampling Search in the Photostore can be replaced with a direct track access, thereby reducing overall access time by perhaps as much as 40 per cent.

These four intended applications for the rapid-search subtable would exploit this technique to effect a significant decrease in processing time and an increase in flexibility and reliability. From presently available information, these tables may well range in size from 25 to 50 thousand characters. This additional capacity of core memory, and the attendant index registers required, are by no means insignificant cost items; however, the production cost with these features will be considerably lower than that anticipated without rapid subtable search.

2.3.3 Production System Performance Estimates

In evaluating the merits of various potential improvements in data processing techniques which might be incorporated into the AN/GSQ-16 Language Processor, a definite need became apparent for a detailed estimate of system performance as measured by processing time. Accordingly, the multiple pass Russian-to-English process was chosen as typical of future operations, and the Mark II organization was assumed as operational with the rapid subtable search (high-speed table search) in core memory. The following charts, processing conditions, and resultant operational performance were derived from this study.

2.3.3.1 Main Dictionary Storage Requirements and Average Entry Lengths (see Tables 2-1, 2-2, and 2-3).

a. Search Input State-Conversion of source language to intermediate language:
   *200,000 entries at 20 char./entry = 4,000,000 char.

b. 16 Intermediate Passes (0 through 15):
   14,860 entries at 36 char./entry = 533,00 char.

c. Final Pass - Conversion of intermediate words into English:
   *300,000 entries at 35 char./entry = 10,500,000 char.

*Best realistic estimate available on operational Russian multipass dictionary.

This represents a total of 15,033,000 characters at 7 bits per character, or a total of 105,200,000 bits.

2.3.3.2 Table-Search Parameters. The statistics on Russian-English Multipass translation for a random sample, 20-word sentence are given in Tables 2-1, 2-2, and 2-3. Regardless of the storage media or method of search, the number of entries that must be compared during an "in-line" (entry) search is determined by:

$$Nx + 1/2\ N$$

where N = number of entries per block (determined by method of "neighborhood" or binary search employed), and Nx = the average number of entries between the desired entry and the place at which ΣI signal changes from "greater than" to "less than." Nx will be assumed to be 15 entries for all searches.

Table 2-1, Search Input State Statistics for Multipass Translation of Russian Sentence to English

Search Input State - Conversion of Source-Language Words Into Intermediate Words:

a) Sample 20-word sentence required a total of 40 searches.

b) With a selected core memory table consisting of items described in c), there would be 30 core table searches and 10 photostore searches.

c) Search Input State core memory table:

10 Searches	1) Punctuation and format control:	500	entries at	15	char/entry		=	7,500 char.
	2) Transliteration :	90	entries at	15	"	"	=	1,350
10 Searches	3) Endings :	1600	" "	25	"	"	=	40,000
10 Searches	4) Common words covering 50% of contextual probability :	500	" "	20	"	"	=	10,000
					Total		=	58,850 char.

Table 2-2, Intermediate Passes - Sentence Analysis Statistics

Pass No.	Total No. of Entries, Average No. of Characters Entry	Total Table Size in Characters	No. of "Control" Entries, Avg. No. of Characters Per Entry	No. of Searches on "Control" Entries	No. of "Additional" Entries	No. of Searches on "Additional" Entries	No. of Main Dictionary Searches	No. of Core Table Searches
0	100 Entries, 20 Char. /Entry	2,000	100 Entries, 20 Char. /Entry	22	—	—	—	22
1	500, 35	17,500	70, 20	47	430	6	—	53
2	240, 40	9,600	240, 40	29	—	—	—	29
3	1060, 35	37,000	50, 25	25	1010	1	1	25
4	400, 35	14,000	400, 35	29	—	—	—	29
5	180, 35	6,300	180, 35	27	—	—	—	27
6	300, 40	12,000	300, 40	24	—	—	—	24
7	Undefined	33,300						32
8	700, 40	28,000	50, 25	22	650	1	—	23
9	3000, 45	135,000	500, 25	44	2500	2	2	44
10	360, 25	9,000	75, 20	31	285	2	—	33
11	Undefined	33,300						32
12	1600, 30	48,000	250, 20	33	1350	8	8	33
13	Undefined	33,300						32
14	Undefined	33,300						32
15	2700, 30	81,000	300, 20	40	2400	1	1	40

Note 1: Data for the four linguistically undefined passes are based on the average statistics of the twelve (12) defined passes.

Note 2: For a 20-word sentence, a total of 522 searches is made for passes 0 through 15. For the AN/GSQ-16 System the breakdown of searches shown in the last two columns is for an 8,000-word (36 bits/word) memory, with core table size restricted to 35,000 characters, providing for processing two sentences, of 150 words each, at a time. In this case there are 510 core table searches and 12 Photostore searches.

Note 3: With a 16,000-word memory, there would be 84,000 character addresses for core tables. In this case there would be 520 core table searches and 2 Photostore searches.

Note 4: Total storage requirement for passes 0 through 15: 533,000 characters.

Table 2-3. Final-Pass Statistics

Conversion of Intermediate words into English:

a) Sample 20-word sentence required a total of 89 searches.

b) With a selected core memory table consisting of items described below, there would be 79 core table searches and 10 Photostore searches.

c) Final-Pass core memory table:

69 Searches	1) Punctuation and format:	500 entries at 25 char./entry	=	12,500 char.
	2) Transliteration, word order, word insertion, English suffixes:	500 entries at 23 char./entry	=	11,500
10 Searches	3) Common words - 50% probability:	500 entries at 35 char./entry	=	17,500
		TOTAL	=	41,500 char.

The average number of characters matched per entry compared, and the average length of function readout is shown in Table 2-4.

Table 2-4, Entry statistics for Search Operations

Table Type	Total Entry Length	Average Number Matching Characters Neighborhood Search	Average Number Matching Characters In-Line Search	Function Readout
Search-Input Pass	20 char.	3 char.	6 char.	12 char.
Intermediate Passes	36 char.	4 char.	9 char.	20 char.
Final Pass	35 char.	4 char.	9 char.	20 char.

Note: All these figures include $\alpha_1 \alpha_2$.

2.3.3.3 Unit Processing Record The expectation for sentence length of technical Russian text is 21 words. Therefore, in order to obtain realistic estimates, throughput calculations should be based on processing of 20-word sentences; two 20-word sentences processed as a unit record.

2.3.3.4 Estimation of AN/GSQ-16 Performance.

a. Operating performance will be calculated under the conditions specified in paragraphs 2.3.3.1 to 2.3.3.3. Photostore parameters, considered necessary and realizable, are:

   120-million bit capacity

   3.5-megacycle bit reading rate

   20-ms average random access time

   11-ms average sequential access time

b. Core memory with 2.4 μ sec cycle time assumed. Throughput rates will be calculated for two sizes of core memory capacity:

1. An 8,000-word (6 characters) memory, allowing approximately 35,000 characters for table storage.
2. A 16,000-word (6 characters) memory, allowing approximately 84,000 characters for table storage.

c. Input/Output via magnetic tape is assumed.

d. Because of core memory limitations, Pass 9 must be separated into "control" entries which will be searched in core tables and "additional" entries which will be searched in the Photostore. This is also true of Passes 3, 12, and 15 with an 8,000-word memory capacity. However, even with a 16,000-word memory, it would be desirable to split up Passes 3, 8, and 15 because the readout times for the additional tables is larger than the Photostore search times required if these tables are split.

e. Core memory table search time depends on several factors -- method used, number of entries, average entry length, Nx, and hardware implementation. The following estimates are based on Nx = 15, a block-stepping followed by entry search technique, and a conservative hardware implementation which saves hardware by requiring that every alternate memory cycle be used for character comparison. Also, these estimates are based on a 2.4 μ sec memory cycle.

1) Search Input Table:

8K:	1090 entries (18,850 char.):	1.4 ms/search
16K:	2690 entries (58,850 char.):	2.0 ms/search

2) Pass	0:	100 entries,	0.80 msec/search
3) Pass	1:	500 entries,	1.20 " "
4) Pass	2:	240 "	1.10 " "

(Split)	5)	Pass 3:	50 entries	0.70 msec/search	
	6)	Pass 4:	400 "	1.10 "	"
	7)	Pass 5:	180 "	1.00 "	"
	8)	Pass 6:	300 "	1.10 "	"
	9)	Pass 7:	930 "	1.50 "	"
(Split)	10)	Pass 8:	50 "	0.70 "	"
(Split)	11)	Pass 9:	500 "	1.20 "	"
	12)	Pass 10:	360 "	1.00 "	"
	13)	Pass 11:	930 "	1.50 "	"
(Split for 8K)	14)	Pass 12:	8K: 250 entries	1.00 "	"
			16K: 1600 entries	1.60 "	"
	15)	Pass 13:	930 "	1.50 "	"
	16)	Pass 14:	930 "	1.50 "	"
(Split)	17)	Pass 15:	300 "	1.00 "	"

18) Final Pass Table:

8K: 1000 entries (24,000 char.): 1.4 ms/search
16K: 1500 entries (41,500 char.): 1.6 ms/search

f. Throughput Rate Estimate:

1) Search Input Pass: (20-word sentence)

	8K:	20 core table searches @ 1.4 ms =	28 ms
		20 Photostore searches @ 20 ms =	400 ms
or	16K:	30 core table searches @ 2.0 ms =	60 ms
		10 Photostore searches @ 20 ms =	200 ms

8K: 428 ms, 16K: 260 ms

2) Intermediate Passes: (20-word sentence)

Pass 0:	22 core table searches	@ 0.8 ms	= 17.6 ms
Pass 1:	53 " " "	@ 1.2 "	= 63.5 "
Pass 2:	29 " " "	@ 1.1 "	= 31.9 "
Pass 3:	25 " " "	@ 0.7 "	= 17.5 "
	1 Photostore search	@ 11 "	= 11.0 "

Pass 4:	29 core table searches	@ 1.1 ms	=	31.9 ms
Pass 5:	27 " " "	@ 1.0 "	=	27.0 "
Pass 6:	24 " " "	@ 1.1 "	=	26.4 "
Pass 7:	32 " " "	@ 1.5 "	=	48.0 "
Pass 8:	22 " " "	@ 0.7 "	=	15.4 "
	1 Photostore search	@ 11 "	=	11.0 "
Pass 9:	44 core table searches	@1.2 "	=	52.8 "
	2 Photostore searches	@11 "	=	22.0 "
Pass 10:	33 core table searches	@1.0 "	=	33.0 "
Pass 11:	32 " " "	@ 1.5 "	=	48.0 "
Pass 12:	8K: 33 core table srch.	@ 1.0 ms	=	33.0 "
	8 Photostore srch.	@ 11 "	=	88.0 "
or	16K: 41 core table srch.	@ 1.6 "	=	65.6 "
	8K: 196.0 ms, 16K: 65.6 ms			
Pass 13:	32 core table searches	@ 1.5 ms	=	48.0 "
Pass 14:	32 " " "	@ 1.5 "	=	48.0 "
Pass 15:	40 " " "	@1.0 "	=	40.0 "
	1 Photostore search	@ 11 "	=	11.0 "

TOTAL: 8K Memory: 719 ms

16K Memory: 661 ms

3) Final Pass: (20-word sentence)

8K:	69 core table searches	@ 1.4 ms	=	96.7 ms
	20 Photostore searches	@ 20 "	=	400.0 "
or 16K:	79 core table searches	@ 1.6 "	=	126.5 "
	10 Photostore searches	@ 20 "	=	200.0 "

8K: 497 ms, 16K: 327 ms

4) Table Accession and Readout:

For two 20-word sentences:

Accession:	8K:	9 Photostore searches @ 11 ms	=	99 ms
or	16K:	3 Photostore searches @ 11 ms	=	33 ms
Readout:				
	8K:	282,450 char. @ 2μ sec	=	565 ms
or	16K:	382,950 " " " "	=	766 ms

For two 20-word sentences:	8K memory:	664 ms
or	16K memory:	799 ms
Per 20-word sentence:	8K memory:	332 ms
or	16K memory:	400 ms

5) Input/Output Time:

729II Magnetic tape, low density: 15,500 char/sec

Two 20-word sentences @ 6 char/word = 240 char

Tape reading $\frac{240}{15,500}$ = 16.0 ms

Start time = 10.0 "

Total I/O time, two 20-word sentences = 26.0 ms

per 20-word sentence: 13 ms

6) Summary: 20 Word Sentence

	8K Memory	16K Memory
Search Input Pass:	428 ms	260 ms
Intermediate Passes:	719	661
Final Pass:	497	327
Table Accession:	50	17
Table Readout:	282	383
I/0 Time:	13	13
	1989 ms	1661 ms
Processing Rate	10 words/sec	12 words/sec

2.3.3.5 Conclusions. These figures reveal that about 25 percent of the processing time is expended in transferring data in and out of the system. Both I/O and table accession and transfer may be accomplished concurrent with internal operations if limited multiplexing is adopted. The resultant production rate would increase 25 percent from 12 to 15 words/second.

Further increases may be realized by structuring the lexicon entries for cascaded entry searches with an attendant reduction in entry size and accession time. Here again a 25 per cent reduction in processing time may be anticipated so that an operational version of the Mark II system could realistically achieve a 20 word/second processing rate. The system processing improvements required to achieve this production rate are realizable with state-of-the-art computer components, such as the two-microsecond core memory.

Further increases in processing rate would require the utilization of developmental higher-performance system components, such as a one-half to one-tenth microsecond core memory cycle, and of a system organization with a much greater degree of parallel operation. By paralleling Photostore and core memory table processing with the faster memory cycle times mentioned, it will be possible to extend the processing rate well up into the 20 to 100-word per second range.

Section 3: CONCLUSION AND RECOMMENDATIONS

3.1 General

The accomplishments reported in the preceding section indicate that significant progress has been made in improving the performance of the language processor. These accomplishments vary in nature from linguistic research to machine organization and in degree from completed tasks which can be operationally demonstrated to studies and designs which will markedly influence future performance of the AN/GSQ-16 complex.

Specifically, an operational lexicon has been completed which employs the bidirectional single-pass translation technique for the translation of Russian to English. This lexicon has been demonstrated to be adequate for immediate requirements. Continued research and development has been devoted to the newer and more powerful lexical processes embodying grammatical analysis and sentence structure determination. Initial studies have also been completed for an improved system organization.

These accomplishments are another significant step in the evolutionary program in machine translation jointly conducted by the Rome Air Development Center and by IBM Research. As such, they quite naturally suggest the next step in this program, that is, the phase of activity which will lead most directly to an operational complex capable of producing the best possible translations.

The areas recommended for immediate consideration also encompass applied and theoretical linguistics as well as system and Photostore improvement studies. These areas are briefly described below and are more thoroughly discussed in the subsections which follow:

a) continued improvement in the RMD and the operational bidirectional translation dictionary;

b) a complete transformation of the RMD into multipass format, generation of new grammar tables for analysis, and initiation of evaluation testing of the combined multipass lexicon against random text;

c) continued basic linguistic research into specific machine translation problems, data compilation for the grammatical feature of government, and Russian grammar;

d) continued studies and logical designs for improved search algorithms and processing techniques for increased production efficiency;

e) further development of the Photostore to improve the throughput of the entire language processing system.

3.2 Applied Linguistics

The development of the bidirectional single-pass translation system proved the usefulness of the approach to Russian-English mechanical translation. Great quantities of randomly selected Russian texts were translated by this program and the translations were considered useful and acceptable by the readers. It is recommended that further effort be spent on the development of this system - both in broadening the analysis capabilities of the bidirectional program and in expanding the lexical contents of the RMD used in conjunction with this program. Also, it is recommended that the RMD be reorganized into microglossaries along the lines discussed in this report.

The work on the multipass program saw further development of the grammatical capabilities of the various passes, embodying almost all of the originally proposed linguistic rules. In addition to this work, work was performed on a program for automatic conversion of RMD entries to search input state, final-pass, and backup multipass entries. A new file, called RMDV, was automatically constructed, to be later acted upon by a relatively simple computer program, yielding the desired multipass entries. It is proposed that this will serve the multipass needs in the way the RMD serves those of the bidirectional program - i.e., that all additions to the RMD be converted into the RMDV format, to be then converted into the multipass entries by the standard conversion program.

Further support of these efforts would yield a fully operative multipass translation system capable of translating random text. It is foreseen that the testing out of this program could be commenced around the middle of the present year.

3.3 Linguistic Research

Research is progressing in two vital and specifically MT areas of interest - a so-called housekeeping program and search strategies. Work in the former area has developed a specific program for which the final details are in process of elaboration. The next step will involve programming, already the subject of many discussions. Extensive testing will then be required in conjunction with an optimal search strategy or even with several search strategies. As for the second area of interest,search strategies, no definitive scheme has been worked out. But a useful critique of existing search strategies has been written and included in this report. This critique, as it were, sets the stage for the development of the kind of search strategy that will permit the most efficient manipulation of grammatical rules and recognition of sentence structure. The next task in this area would seem to be a specification of the essential elements of a search strategy and their relative evaluation as a basis for the construction and testing of the best type of search strategy.

An ambitious program for the massive compilation of grammatical data for the Russian language has been outlined. This program is built around the grammatical feature of government and, as such, is primarily concerned with verbal government. This kind of information is of prime importance for any MT system involving Russian and for any program of grammatical research on Russian. This information represents basic data for any Russian-English MT system and basic input for any research program on Russian grammar. All types of pertinent or even supposedly pertinent information have been classified and incorporated into tables for use by prospective data compilers. The tables attempt to reduce the work of compiling the masses of data to a relatively small number of controlled steps. The tables are next to be tested on an adequate number of lexical items before they are presented to a group of grammarian-compilers.

An extensive and long-range program for grammatical research in Russian has been instituted in accordance with the most powerful and most interesting concept of grammatical structure. More specifically, this program aims at a rigorous formulation of grammatical rules within the framework of a generative grammar. This work will not only contribute to knowledge of Russian grammar within the scholarly community but also will generate well considered data for Russian-English MT and will result in a better understanding of grammatical rules so that more sophisticated MT systems can be constructed. Research in this area of IBM will continue to concentrate on constituent structure rules.

as a basis for the construction and testing of the best type of search strategy.

An ambitious program for the massive compilation of grammatical data for the Russian language has been outlined. This program is built around the grammatical feature of government and, as such, is primarily concerned with verbal government. This kind of information is of prime importance for any MT system involving Russian and for any program of grammatical research on Russian. This information represents basic data for any Russian-English MT system and basic input for any research program on Russian grammar. All types of pertinent or even supposedly pertinent information have been classified and incorporated into tables for use by prospective data compilers. The tables attempt to reduce the work of compiling the masses of data to a relatively small number of controlled steps. The tables are next to be tested on an adequate number of lexical items before they are presented to a group of grammarian-compilers.

An extensive and long-range program for grammatical research in Russian has been instituted in accordance with the most powerful and most interesting concept of grammatical structure. More specifically, this program aims at a rigorous formulation of grammatical rules within the framework of a generative grammar. This work will not only contribute to knowledge of Russian grammar within the scholarly community but also will generate well considered data for Russian-English MT and will result in a better understanding of grammatical rules so that more sophisticated MT systems can be constructed. Research in this area of IBM will continue to concentrate on constituent structure rules.

3.4 System Organization

The essential operating characteristics of the Mark II AN/GSQ-16 language processing system have been carefully reviewed during the past year, with the objective of identifying those areas where changes in operating technique could materially increase the utility of its general-purpose and universal table process and improve the effectiveness, efficiency, and reliability of data processing.

A thorough study has been conducted of the optimum method for performing rapid searches of integrally addressed lexical subtables, or entry sets, stored in a character-addressable core memory (high-speed table search). The application of such subtables to common words, addenda, errata, and indices has been explored.

As a result of these studies, some specific modifications to the search and analysis logic have been formulated. These modifications are designed to implement the rapid subtable search routine, which will improve the translation rate about one to two orders of magnitude. In the course of this investigation several novel processing techniques were conceived which could effect considerable improvements in the logic routines. It is recommended that these improvements be studied in detail to determine, first, their feasibility and, second, their logical design implementation. The improved processing features fall into three categories which are briefly identified below.

3.4.1 Search Algorithms

In addition to the rapid subtable search routine, several search routine modifications are apparent which improve processing capabilities. Cascaded searches of logically structured tables, a trace routine for lexical program diagnosis and data analysis, search repetition on detected errors, and a logical comparison capability may add significantly to system performance.

3.4.2 Data Processing

Natural language string processing can be considerably improved by providing a means of tagging this string directly with specific data obtained during the process. In multiple pass analysis, valuable additions in programming capabilities may be obtained by providing a means of adding, deleting, or rearranging the intermediate lexical words. Finally, special tables which are selected by, or developed from, the input text may prove quite effective in processing proper names, unique errors, and unusual linguistic structures or usages.

3.4.3 System Optimization for Processing Efficiency

The AN/GSQ-16 system organization is being developed to the point where an operating configuration based upon a production environment should be formulated. Efficient procedures can be developed to overlap many processes now performed sequentially. Other procedures can be employed in the language processor to effect a significant generality in application and a wide compatibility with other data processing.

3.5 Photostore Improvements

The experience gained from the existing Photostore in its regular use as part of the language translation complex has suggested certain improvements that could be made to the unit, which would improve the throughput of the entire AN/GSQ-16 system.

These improvements relate to the capacity and reading rate of the Photostore and its reliability. An effort is being made to create a cleaner environment in the Photostore, which would have a definite bearing on its reliability. The improvements recommended are described a little more fully in the following paragraphs.

3.5.1 Disk Capacity

The capacity of the Photostore disk now in use is 60 million bits. The disk uses a 0.360-inch annulus, which represents only a fraction of the available emulsion area. Consideration is being given to a mechanical effort directed toward improving the actuator so that it is capable of a 1-inch throw, which would increase the available disk storage area by a factor of 2.5. This change in the throw of the actuator would make possible the storage of up to 145,000,000 bits at the present density.

3.5.2 Increased Reading Rate

The reading rate of the Photostore has been increased during the past year from 1 mc to 1.55 mc. With the type of circuits now employed in the disk reader and Mark II system, it is believed that a reading rate of at least 2 mc will be achieved in the coming year.

3.5.3 Reliability

Investigation of present disk-making techniques indicate that the reliability limit of existing equipment is rapidly being approached. An alternate technique for disk-making, which has both the advantages of fast writing and precision control of mark positioning, has been under study by IBM Research. This technique eliminates the film intermediate by using an electron beam that writes directly on the silver halide emulsion on the disk at a rate of several hundred thousand bits per second. It is expected that this novel technique will play an important role in improving photoscopic disk quality.

3.5.4 Environmental Cleanliness

It has become apparent during the past year that excellent reliability can be achieved only after improved techniques have been developed for storage and transfer of disks and a cleaner environment is maintained in the reader.

It is recommended that an effort be devoted to the development of handling and storage techniques that will prevent contamination of the disks, and the development of a hood for the reader that will prevent the dropping of dirt particles from the reader onto the disk when the disk is not rotating.

3.5.5 Studies of Increased Density

Before increasing the bit density further, it appears desirable to initiate a thorough analysis of the disk-making procedure and the electro-optical portion of the disk reader. The objective of these studies is to recognize and evaluate all pertinent characteristics of the system as they affect system reliability. It is anticipated that these studies would pinpoint any marginal elements in the system, provide guidance in development work, and permit realistic signal-to-noise and, hence, reliability predictions as a function of disk bit density.

APPENDICES

Доклады Академии наук СССР
1957. Том 117, № 2

ФИЗИКА

В. И. БЕСПАЛОВ

К ВОПРОСУ О ФЛУКТУАЦИЯХ ПАРАМЕТРОВ НЕКОТОРЫХ ЛИНЕЙНЫХ СИСТЕМ

(Представлено академиком М. А. Леонтовичем 8 VI 1957)

1. Специфической особенностью задачи о рассеянии волн, распространяющихся в экранированной линии передачи, на имеющихся в линии случайных неоднородностях является то обстоятельство, что вторичные (переизлученные неоднородностями) волны канализируются тем же трактом, что и первичная волна. Если длина линии достаточно велика, то вторичное поле (т. е. амплитуды отраженной волны и волн других типов, возникающих в результате перетрансформации) может оказаться сравнимым с полем падающей волны. По этой причине метод возмущений, используемый обычно в задачах о рассеянии ([1,2]) и не учитывающий вторичного переизлучения рассеянных волн, оказывается недостаточным для решения ряда вопросов о влиянии случайных неоднородностей на характеристики линий передачи.

Аналогичные трудности возникают также при решении ряда других задач, где речь идет о влиянии случайных отклонений параметров на характеристики линейной системы. В качестве примера можно привести: фильтр (или линию задержки), параметры ячеек которого имеют некоторый случайный разброс от номинальных значений; лампу с бегущей волной, в которой использована замедляющая система со случайными нарушениями структуры; колебательный контур, параметры которого меняются скачками случайной величины, и т. п. Решение такого рода задач может быть сведено, при известных допущениях (см., например, ([3])), к исследованию системы линейных разностных уравнений.

$$Y_j(n) = A_{jk}(n)\, Y_k(n-1), \quad j, k = 1, 2, \ldots, L, \tag{1}$$

коэффициенты которых являются случайными функциями n.

Нахождение наиболее полной характеристики случайного процесса $Y_l(n)$ — распределения плотностей вероятности $W(Y_l, n)$ — связано со значительными трудностями и может быть выполнено либо в результате исследования общего решения системы (1), либо путем решения соответствующих дифференциально-разностных уравнений для W. Однако во многих случаях необходимую информацию о случайном процессе дают моменты и функции корреляции.

В настоящей работе приводится общее решение системы (1) и сравнительно простой метод вычисления моментов и функций корреляции. В качестве примера рассмотрена цепочка простейших Г-образных четырехполюсников и колебательный контур с флуктуирующими параметрами.

2. Для отыскания решения системы разностных линейных уравнений с переменными коэффициентами может быть использован метод последовательных приближений. В отличие от дифференциальных уравнений, где в каждом отдельном случае необходимо доказывать сходимость ряда приближений, здесь всегда можно выбрать нулевое приближение таким образом, что на любом ограниченном интервале конечное число последовательных приближений приводит к точному решению. При этом, очевидно, первоначальное предположение о малости возмущения коэффициентов становится излишним.

Запишем коэффициенты уравнений (1) в виде

$$A_{jk}(n) = A_{jk}^{0} + \mu a_{jk}(n). \tag{2}$$

Решение системы (1), удовлетворяющее начальным условиям

$$Y_j(n)|_{n=0} = C_j, \tag{3}$$

будем искать в виде ряда по степеням μ

$$Y_j(n) = \sum_{s=0}^{\infty} \mu^s Y_j^{(s)}(n). \tag{4}$$

Тогда, подставляя (4) в (1) и группируя члены с одинаковыми степенями μ, получим

$$Y_j^s(n) = A_{jk}^{0} Y_k^{(s)}(n-1) + a_{jk}(n) Y_k^{(s-1)}(n-1). \tag{5}$$

Если потребовать, чтобы $Y_j^{(0)}(n)$ удовлетворяли начальным условиям (3), то, как нетрудно видеть, ряд обрывается при $s = n$, так как все $Y_j^{(s)}$ при $s > n$ тождественно обращаются в нуль. Действительно, $Y^{(s)}(0) \equiv 0$ при $s > 0$ по выбору нулевого приближения. Из (5) очевидно, что $Y_j^{(s)}(n) = 0$, если $Y_j^{(s-1)}(n-1) = 0$ и $Y^{(s)}(n-1) = 0$. Согласно методу индукции отсюда следует, что $Y^{(s)}(n) \equiv 0$ при $s > n$.
Выбирая соответствующим образом C_j, можно удовлетворить любым граничным условиям (заданным не обязательно при $n = 0$).

Например, общее решение уравнения второго порядка с одним переменным коэффициентом *

$$Y(n+1) - [\Phi_0 + P(n)]\, Y(n) + Y(n-1) = 0, \tag{6}$$

полученное методом последовательных приближений, имеет вид

$$Y(n) = Ae^{j\varphi_0 n} \times$$

$$\left\{1 + \sum_{s=1}^{n-1} (2j\sin\varphi_0)^{-s} \sum_{f_s=1}^{n-1} \sum_{f_{s-1}=1}^{f_s-1} \ldots \sum_{f_1=1}^{f_2-1} \prod_{m=1}^{s} P_{f_m}\left[e^{j2\varphi_0(f_m-f_{m+1})} - 1\right]\right\} +$$

$$+ Be^{-j\varphi_0 n} \times$$

$$\left\{1 + \sum_{s=1}^{n-1} (2j\sin\varphi_0)^{-s} \sum_{f_s=1}^{n-1} \sum_{f_{s-1}=1}^{f_s-1} \ldots \sum_{f_1=1}^{f_2-1} \prod_{m=1}^{s} P_{f_m}\left[1 - e^{-j2\varphi_0(f_m-f_{m+1})}\right]\right\}, \tag{7}$$

где A и B — произвольные постоянные и $\varphi_0 = \arccos(\Phi_0/2)$.

* Такого вида уравнение получается при исследовании цепочки четырехполюсников или с переменным последовательным сопротивлением или с переменной шунтирующей емкостью.

Если величины $P(n)$ относительно малы ($\mu \ll 1$), то во многих случаях можно ограничиться первым [3] или вторым приближением.

Повышение порядка разностного уравнения (системы) и усложнение коэффициентов приведет к дальнейшему увеличению громоздкости решения, что ограничивает возможность его применения при изучении случайных процессов. Вместе с тем, если число ячеек мало (например, в фильтрах), общее решение типа (7) может быть с успехом использовано для исследования возмущений, обусловленных случайными изменениями параметров ячеек.

3. Средние * характеристики величин, описываемых (1), можно найти достаточно просто в том случае, когда процесс $Y^j(n)$ является простой цепью Маркова. Последнее имеет место, если:

а) случайные функции $A_{jk}(n)$ некоррелированны **, т. е.

$$\overline{A_{jk}(n)\,A_{lp}(m)} = \overline{A_{jk}(n)}\ \overline{A_{lp}(m)}; \tag{8}$$

б) граничные (начальные) условия заданы при одном и том же значении n, например при $n=0$,

$$Y_j(n)|_{n=0} = Y_j^0. \tag{9}$$

При этом уравнения и начальные условия для $\overline{Y_j(n)}$ получаются сразу же путем усреднения (1) и (9), т. е.

$$\overline{Y_j(n)} = \overline{A_{jk}(n)}\ \overline{Y_k(n-1)}; \quad \overline{Y_j(n)}|_{n=0} = Y_j^0. \tag{10}$$

Таким образом, средние значения $\overline{Y_j(n)}$ в цепочках со случайными параметрами при выполнении условий (8) и (9) распределены так же, как и величины $Y_j(n)$ в цепочке со средними значениями параметров $\overline{A_{jk}(n)}$ и усредненными начальными условиями ***.

При тех же предположениях (8) и (9) система уравнений для средних значений произведений $\overline{Y_j(n)\,Y_k^*(n)} = \xi_{jk}(n, 0)$ получается после усреднений произведений уравнений (соответственно, правых и левых частей) системы (1) на уравнения комплексно сопряженной системы ****. В результате получим L^2 уравнений

$$\xi_{jk}(n, 0) = \overline{A_{jl}(n)\,A_{kp}^*(n)}\,\xi_{lp}(n-1, 0). \tag{11}$$

* Здесь и в дальнейшем имеется в виду усреднение по ансамблю.

** При наличии корреляции в конечной области, т. е. если (8) выполняется только при $|m-n| > \nu \neq 0$, процесс $Y_j(n)$ является сложной цепью Маркова. Последняя может быть сведена к простой цепи [4], однако уравнения значительно усложнятся.

*** Отметим, что, вообще говоря, среднее значение напряжения и тока в ансамбле цепочек четырехполюсников без потерь может возрастать или убывать по экспоненциальному закону.

**** Уравнения для моментов второго порядка (или других средних величин) можно получить и другими путями. В частности, можно воспользоваться дифференциально-разностным уравнением (которое здесь будет бесконечного порядка) для вероятностей перехода или же перейти от разностного уравнения (1) к суммовому (аналог интегрального) и использовать метод, примененный в работе [5] для исследования интегрального уравнения со случайным ядром. Используемый здесь метод приводит к нужному результату более коротким путем.

Если элементы матрицы коэффициентов (11) не зависят от n, то, полагая***** $\xi_{jk}(n, 0) = \xi_{jk}(0, 0)\alpha^n$, из условия существования нетривиального решения найдем

$$\det \left| \overline{A_{jl}(n) A^*_{kp}(n)} - \alpha\delta_{jl}\delta_{kp} \right| = 0, \tag{12}$$

где δ_{qf} — символ Кронекера. Уравнение (12) определяет L^2 значений α и, следовательно, L^2 линейно независимых решений $\xi_{jk}(n, 0)$, с помощью которых можно удовлетворить начальным условиям

$$\xi_{jk}(n, 0)|_{n=0} = \overline{Y^0_j Y^{0*}_k}. \tag{13}$$

Умножая (1) на $Y^*_k(n-m)$ и усредняя, для функций корреляций $\xi_{jk}(n, m) = \overline{Y_j(n) Y^*_k(n-m)}$ получим L (соответственно индексу k, принимающему значения от 1 до L) систем, каждая из которых состоит из L разностных уравнений

$$\xi_{jk}(n, m) = \overline{A_{jl}(n)}\, \xi_{lk}(n-1, m-1). \tag{14}$$

Решение систем (14) при известных $\xi_{jk}(n, 0)$ находится просто. Аналогично можно вычислить и моменты более высоких порядков.

4. Распределение среднего квадрата модуля напряжения в цепочке простейших Г-образных четырехполюсников с флуктуирующими последовательными сопротивлениями ($X_n = X_0\{1 + P(n)\}$), полученное приведенным методом в предположении, что $P(n)P(m) = p^2\delta_{nm}$, $p^2 \ll 1$, и нагрузка на выходе согласована, имеет вид:

$$\overline{|V(n)|^2} \simeq |V_0^2| \left\{ \left(1 - p^2 \operatorname{tg}^2 \frac{\varphi}{2}\right) \exp\left(2p^2 n \operatorname{tg}^2 \frac{\varphi}{2}\right) - \right.$$
$$\left. - p^2 \operatorname{tg}^2 \frac{\varphi}{2} \frac{\sin(2n-1)\varphi}{\sin\varphi} \cdot \exp\left(-p^2 n \operatorname{tg}^2 \frac{\varphi}{2}\right) \right\}, \tag{15}$$

где φ — сдвиг фазы на один четырехполюсник. Из (15) следует, что при больших n величина $\overline{|V^2(n)|}$ возрастает по экспоненциальному закону. Аналогично средний квадрат модуля тока в колебательном контуре с емкостью флуктуирующей по закону $C(t) = C_0\{1 + P(n)\}^{-2}$, где t — время и n — целая часть t/τ, описывается формулой

$$\overline{|I^2(n)|} \simeq I_0^2 \left\{ \left[1 - p^2 \left(\sin^2\omega\tau - \frac{\omega\tau \sin 2\omega\tau}{2\sin^2\omega\tau}\right)\right] \exp(2p^2 n \sin^2\omega\tau) - \right.$$
$$\left. - p^2 \left[\frac{\sin^2\omega\tau \cdot \sin 2(n-1)\omega\tau}{\sin 2\omega\tau} + \frac{\omega\tau \cos(2n+1)\omega\tau}{\sin\omega\tau}\right] \exp(-p^2 n \sin^2\omega\tau) \right\}, \tag{16}$$

где ω — резонансная частота идеального контура; τ — время, через которое изменяется емкость.

Уравнение (16) показывает, что из-за флуктуаций емкости запасенная энергия в ансамбле контуров со временем возрастает, причем при $t \to \infty$ максимумы $\frac{1}{2}L\overline{|I^2(t, \tau)|}\Big|_{t=\mathrm{const}}$ наблюдаются при τ, удовлетворяющих уравнению $2\omega\tau \operatorname{ctg}\omega\tau = 1$. Последнее совпадает с условием максимума среднего квадрата плотности амплитуды спектра функции $C(t)$ на одной из частот, соответствующих параметрическому резонансу ([6]). Полученные результаты нетрудно обобщить на системы со случайно изменяющимися интервалами τ.

***** Если статистические свойства A_{jk} зависят от n (например, на заданную зависимость $A_{jk}(n)$ накладываются случайные возмущения), то $\overline{A_{jl}(n) A^*_{kp}(n)} \neq \mathrm{const}$ и решение системы (11) записывается в виде конечного числа многократных сумм (п. 2). Трудность исследования решения в каждом конкретном случае будет определяться характером зависимости A_{jk} от n.

5. Приведенный метод расчета усредненных величин может быть использован не только при изучении систем со ступенчатым изменением параметров, но также и систем с непрерывными флуктуациями, если форму последних (в пространстве или времени) можно с определенной степенью точности считать повторяющейся. Таким методом, например, можно оценить влияние неоднородностей в цилиндрических линиях передач ([7]).

В заключение автор выражает признательность А. В. Гапонову за советы при выполнении работы.

Научно-исследовательский радиофизический институт
при Горьковском государственном университете
им. Н. И. Лобачевского

Поступило
3 V 1957

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] С. М. Рытов, Изв. АН СССР, сер. физ., № 2, 223 (1937). [2] H. Booker, W. Gordon, Proc. Inst. Rad. Eng., 38, 4, 401 (1950). [3] В. И. Беспалов, А. В. Гапонов, Радиотехника и электроника, 1, № 6, 772 (1956). [4] Т. А. Сарымсаков Основы теории процессов Маркова, М., 1954., 1954. [5] A. Rosenbloom, J. Heilfron, D. L. Trautman, Inst. Rad. Eng. Convent. Rec., 3, 4, 106 (1955). [6] А. А. Андронов, М. А. Леонтович, Ж.Р.Ф.-Х.О., ч. физ., 59, № 5—6, 429 (1927). [7] M. D. Didlandis, H. Kaden, Elektrische Nachrihten-Technik. 14, 1, 13 (1937); P. Mertz, W. Pfleger, Bell. Syst. Techn. J., 16, 541 (1937).

Tape No. 2129 R-75
Page No. 209-212 Article No. 1
Bidirectional Single-Pass Translation

Reports of Academy of sciences of/by

USSR 1957. That/volume 117. No. 2

Physicist

V. I. Bespalov

Concerning the Question of Fluctuations of Parameters of Certain Linear Systems

(Presented by academician M. A. Leontovichem 8 VI 1957)

1. Specific peculiarity of problem concerning diffusion of wave, spread in shielded transmission line, on available in line of accidental heterogeneities is that circumstance that secondary (overradiated heterogenities) wave are sewered the same channel that and primary wave. If length of line sufficiently great, that secondary field (i amplitude of/by reflected wave and waves of other types, appearing as a result of *peretransformatsii*) can appear comparable with field of/by falling wave. By this cause perturbation theory, utilized usually in problems concerning diffusion ([1], [2]) and not considering secondary overradiation of absent-minded/scattered. wave, appears insufficient for decision/solution of series/row of questions about influence of accidental heterogeneities on characteristic of transmission line.

Analogous difficulties appear also at/during decision/solution of series/row of other problem, where it is a question concerning influence of accidental deflections of parameters on characteristic of/by linear system. As example it is possible bring: filter (or delay line), parameters of cells of that have certain accidental scattering from nominal values; tube/lamp with traveling wave, in which used delaying system with accidental disturbances of structure; oscillation circuit, parameters of that change races of random variable, and others Decision/solution of such kind of problems can be reduced, at/during known assumptions (see, for example, [3]), to research of system of linear difference equations.

$$Y_j(n) = A_{jk}(n)\,Y_k(n-1), \quad j,\ k = 1,\ 2,\ \ldots,\ L, \tag{1}$$

coefficients of that are function of a randoms n. the most full characteristic of accidental process Y (n) - distribution probability density W (Y_j, n) - combined with significant difficulties and can be carried out or as a result of research of general solution of system (1), or by means of decision/solution of corresponding differential-difference equations for W. However in many cases necessary information concerning accidental process quarter-deck moments and function of correlation.

In this work is brought/listed general/more general solution of system (1) and comparatively idle time/simple method of calculation of moments and functions of correlation. As example considered iterated network of simplest Γ-graphic quadripoles and oscillation circuit with fluctuating parameters.

2. For detecting of solution of system difference linear equations with variable coefficients can be used method of successive approximations. As distinguished from differential equations, where in every separate case *neobkhodimo* prove convergence series/rowa *priblizheniy*, here always it is possible to select zero approximation/approach by thus that on any limited interval final/finite number of consecutive *priblizheniy* brings/lists to accurate decision/solution. With this/besides, obviously, initial assumption against/concerning smallness of disturbance of coefficients turns unnecessary.

Will record coefficients of equations (1) in the form of

$$A_{jk}(n) = A^0_{jk} + \mu a_{jk}(n). \tag{2}$$

Solution of system (1), satisfying initial conditions

$$Y_j(n)|_{n=0} = C_j, \tag{3}$$

will search in ida of series/row along/by degrees μ

$$Y_j(n) = \sum_{s=0}^{\infty} \mu^s Y_j^{(s)}(n). \tag{4}$$

Then, substituting (4) in (1) and grouping members with identical degrees μ, will receive

$$Y_j^s(n) = A_{jk}^0 Y_k^{(s)}(n-1) + a_{jk}(n) Y_k^{(s-1)}(n-1). \tag{5}$$

If demand in order to $Y_j^{(0)}$ (n) satisfied initial conditions (3), that, how/as simply see, series/row is break/stopped at/during s = n, since all $Y_j^{(s)}$ at/during s > n identically are turned in zero. In fact, $Y^{(s)}(0) \equiv 0$ at/during s > 0 along/by selection of zero approximation/approach. Out of (5) it is obvious that $Y_j^{(s)}$ (n) = 0 if $Y_j^{(s-1)}$ (n − 1) = 0 and $Y_k^{(s)}$ (n − 1) = 0. According to method of induction it follows from this that $Y^{(s)}(n) \equiv 0$ at/during s > n.

Selecting in the appropriate way C_j, it is possible to satisfy any boundary conditions (given not certainly/obligatory at/during n = 0).

For example, general/more general solution of equation of the second order with one/only variable coefficient*

$$Y(n+1) - \{\Phi_0 + P(n)\} Y(n) + Y(n-1) = 0. \tag{6}$$

received method of consecutive approximations, has the form of

$$Y(n) = Ae^{j\varphi_0 n} \times$$
$$\left\{1 + \sum_{s=1}^{n-1} (2j\sin\varphi_0)^{-s} \sum_{f_s=1}^{n-1} \sum_{f_{s-1}=1}^{f_s-1} \cdots \sum_{f_1=1}^{f_2-1} \prod_{m=1}^{s} P_{f_m}\left[e^{j2\varphi_0(f_m-f_{m+1})} - 1\right]\right\} +$$
$$+ Be^{-j\varphi_0 n} \times$$
$$\left\{1 + \sum_{s=1}^{n-1} (2j\sin\varphi_0)^{-s} \sum_{f_s=1}^{n-1} \sum_{f_{s-1}=1}^{f_s-1} \cdots \sum_{f_1=1}^{f_2-1} \prod_{m=1}^{s} P_{f_m}\left[1 - e^{-j2\varphi_0(f_m-f_{m+1})}\right]\right\}, \tag{7}$$

*Such form equation happen/obtains at/during research of iterated network of quadripoles or with variable consecutive resistance or with/from variable shunting capacity.

where A and B – arbitrary constant and $\phi_0 = \arc\cos(\Phi_0/2)$.
If magnitude P (n) relatively/concerning small ($\mu << 1$), that in many cases it is possible to be limited first ([3]) or second approximation/approach.

Increase of order of difference equation (system) and complication of coefficients will bring/list to further increase of cumbersomeness of decision/solution, what limits possibility him/his/it/its application during the study of of accidental processes. Together with that, if number of cells little (for example, in filters), general/more general decision/solution of type (7) can be with success used for research of disturbances, stipulated accidental changes of parameters of cells.

3. Average* characteristic of magnitude, described (1), can be found sufficiently simply in that a case, when process Y (n) is idle time/simple chain/circuit *Markova*. Last takes place, if:
but) function of a random A_{jk} (n) uncorrelated**, i

$$\overline{A_{jk}(n)A_{lp}(m)} = \overline{A_{jk}(n)}\ \overline{A_{lp}(m)}; \tag{8}$$

b) threshold (initial) condition are given at/during one and the same value n, for example at/during n = 0

$$Y_j(n)|_{n=0} = Y_j^0. \tag{9}$$

With this/besides equation and initial condition for Y (n) happen/obtain immediately indeed by means of averaging (1) and (9), i

$$\overline{Y_j(n)} = \overline{A_{jk}(n)}\ \overline{Y_k(n-1)}; \quad \overline{Y_j(n)}|_{n=0} = Y_j^0. \tag{10}$$

* Here and in the remainder is in view averaging along/by ensemble.

** At/during presence of correlation in final/finite region, i if (8) is executed only at/during $|m-n| \geq \nu \neq 0$. process Y_j (n) is in a complex way/it is difficulty chain/circuit *Markova*. Last can be reduced to idle time/simple chain/circuit ([4]), however equation significantly will be complicated.

Thus, mean value $\overline{Y_j\ (n)}$ in iterated networks with accidental parameters at/during fulfillment of conditions (8) and (9) distributed the same way, as and magnitude Y_j (n) in iterated network with mean values of parameters $\overline{A_{jk}(n)}$ and neutralized initial conditions***.

With those same assumptions (8) and (9) system of equations for mean values of products $\overline{Y_j(n)\,Y_k^*(n)} = \xi_{jk}(n, 0)$ happen/obtains after averages of products of equations (correspondingly, right and left sides) system (1) on equation complex joint system****. As a result of will receive L^2 equations.

$$\xi_{jk}(n, 0) = \overline{A_{jl}(n)\,A_{kp}^*(n)}\,\xi_{lp}(n-1,\ 0). \tag{11}$$

If elements of matrix of coefficients (11) do not depend on n, that, considering***** ξ_{jk} (n,o) = $\xi_{jk}(0,\ 0)\,\alpha^n$,out of condition of existence of nontrivial decision/solution will discover.

$$\det\left|\overline{A_{jl}(n)\,A_{kp}^*(n)} - \alpha\delta_{jl}\delta_{kp}\right| = 0, \tag{12}$$

where δ_{qf} - kronecker symbol. Equation 12) determines L^2 values α and, consequently L^2 linearly independent solutions ξ_{jk} (n, 0), with the help of that it is possible to satisfy initial conditions

$$\xi_{jk}(n, 0)|_{n=0} = \overline{Y_j^0 Y_k^{0*}}. \tag{13}$$

*** Let us note that, generally speaking, mean value of voltage/effort and current in ensemble of chains quadripoles without rub can increase or diminish along/by exponential law.

**** Equation for moments of the second order (or other average magnitudes) can be obtained and other ways. In particular, it is possible to be used differential-difference equation (that here will be infinite order) for transition probability or indeed cross from difference equation (1) to sum (analog of integral) and use method, applied in work (5) for investigation of integral equation with accidental nucleus. Utilized here method brings/lists to needed result more short way.

***** If statistical properties A_{jk} depend on n (for example, on given dependence A_{jk} (n) are placed accidental disturbances), that $\overline{A_{jl}(n)\,A_{kp}^*}$ (n) $\neq$ const and solution of system (11) is recorded in the form of final/finite number of multiple sums (pood 2) Difficulty of investigation of decision/solution in every concrete case will be determined character of dependence A_{jk} from n.

Multiplying (1) on $Y^*_k(n-m)$ and neutralizing, for functions of correlations $\xi_{jk}(n,m) = \overline{Y_j(n) Y^*_k(n-m)}$ will receive L (correspondingly index k, admitting value from 1 to L) systems, every out of that consists of L difference equations

$$\xi_{jk}(n,m) = A_{jL}(n)\,\xi_{Lk}(n-1,\, m-1). \tag{14}$$

Decision/solution of systems (14) at/during known $\xi_{jk}(n,0)$ is simply. Analogously it is possible to calculate and moments more high orders.

4. Distribution of average square of modulus of voltage/effort in iterated network of simplest Γ-graphic quadripoles with fluctuating consecutive resistances ($X_n = X_0 \{1 + P(n)\}$), received brought method in assumption that $\overline{P(n)\, P(m)} = p^2\delta_{nm}$, $p^2 << 1$. and load on output/exit coordinated, has the form of:

$$\overline{|V(n)|^2} \simeq |V_0^2| \left\{\left(1 - p^2 \operatorname{tg}^2 \frac{\varphi}{2}\right) \exp\left(2p^2 n \operatorname{tg}^2 \frac{\varphi}{2}\right) - p^2 \operatorname{tg}^2 \frac{\varphi}{2} \frac{\sin(2n-1)\varphi}{\sin\varphi} \cdot \exp\left(-p^2 n \operatorname{tg}^2 \frac{\varphi}{2}\right)\right\}, \tag{15}$$

where ϕ – phase shift on one quadripole. Out of (15) one should that at/during large n magnitude $|V^2(n)|$ increases along/by exponential law. Analogously average square of modulus of current in oscillation circuit with capacity of/by fluctuating along/by law $C(t) = C_0 \{1 + P(n)\}^{-1}$, where t – time and n - whole part t/τ, is characterized by formula

$$\overline{|I^2(n)|} \simeq I_0^2 \left\{\left[1 - p^2\left(\sin^2\omega\tau - \frac{\omega\tau \sin 2\omega\tau}{2\sin^2\omega\tau}\right)\right] \exp(2p^2 n \sin^2\omega\tau) - p^2\left[\frac{\sin^3\omega\tau \cdot \sin 2(n-1)\omega\tau}{\sin 2\omega\tau} + \frac{\omega\tau \cos(2n+1)\omega\tau}{\sin\omega\tau}\right] \exp(-p^2 n \sin^2\omega\tau)\right\}, \tag{16}$$

where ω – resonance frequency of ideal circuit/contour; τ – time, through that changes capacity.

Equation 16) shows that because of/from behind fluctuations capacity store energy in ensemble of circuit/contours in time increases, and besides at/during $t \to \infty$ maxima $\overline{1/2\,L\,|\,i^2\,(t,\tau)\,|}_{t=\mathrm{const}}$ are observed at/during τ, satisfying equation $2\omega\tau\,\mathrm{ctg}\,\omega\tau = 1$. Last coincides with condition of maximum of average square of density of amplitude of spectrum of function C (t) on one of frequency, corresponding parametric resonance ([6]). Received results simply generalize on system with/from accidentally changing intervals τ.

5. Brought method of calculation/crew of neutralized magnitudes can be used not only during the study of of systems with step change of parameters, but also and systems with continuous fluctuations, if form of last (in space or time) it is possible with/from definite degree of accuracy count/consider repeated. Such method, for example, it is possible to estimate influence heterogeneities in cylindrical lines of transmissions ([7]).

In conclusion author expresses gratefulness *A. V. Gaponovu* for/after councils at/during fulfillment of work.

Scientific research radio physics institute

at/during Gorki*Gorkovskom* state university

him/it/them. *N. I Lobachevskogo*

Proceeded

3 V 1957

Quoted Literature

[1] *S. M. Rytov*,*Izv.*Academy of Sciences of/by USSR, gray. *fiz.*, No. 2. 22 (1937). [2] H. Booker, W. Gordon, Proc. Inst. Rad. Eng. 38. 4. 401 (1950). [3] *V. I. Bespalov, A. V. Gaponov*, Radio engineering/technician and electronics, 1. No. 6. 772 (1956).*Shch A. Rosenbloom, Zh. Yueilfron, D. L. Trautman, Inst. Rad. Eng.Chonvent. Rech.*, *E*, I,*L*) (1955). [6] *A. A. A ndronov, M. A. Leontovich*, ()*khzh.R,F.*-Chemical reconnaissance.Against/concerning., *ch. fiz.*,59 No. 5–6. 429 (1927). [7] M. D. Didlandis, H. Kaden, Elektrische Nashrihten-Techchnik. 14. 1. 13 (1937); P. Mertiz, W. Pfleger, Bell. Syst. Techn. J. 16.541 (1937).

ЖУРНАЛ
ЭКСПЕРИМЕНТАЛЬНОЙ И ТЕОРЕТИЧЕСКОЙ ФИЗИКИ
ТОМ 42 1962 ВЫП. 3

О ВОЗМОЖНОСТИ УСИЛЕНИЯ УЛЬТРАЗВУКА В ПОЛУМЕТАЛЛАХ В ЭЛЕКТРИЧЕСКОМ ПОЛЕ

Р. Ф. Казаринов, В. Г. Скобов

Взаимодействие звуковой волны с электронами проводимости в кристалле приводит к диссипации звуковой энергии. При этом существенное влияние на прохождение звука может оказать наличие внешнего электрического поля **E**. В работе Хатсона, Мак-Фи и Уайта [1] обнаружено интересное явление: усиление ультразвука в полупроводнике CdS в электрическом поле. В этом кристалле имеется сильное пьезоэлектрическое взаимодействие электронов со звуком. Однако принципиальная возможность усиления звука электронами проводимости в электрическом поле не зависит от конкретного характера взаимодействия. Эффект имеет следующую природу. В отсутствие электрического поля электроны поглощают звуковую энергию Q_0. Это приводит к появлению электронного акустического тока **J**, пропорционального Q_0. Поэтому звуковая энергия, поглощаемая электронами в единицу времени, в линейном по **E** приближении есть $Q = Q_0 + \mathbf{JE}$. Если значение **E** таково, что $Q < 0$, то имеет место усиление звука электронами.

При поглощении звукового кванта скорость электрона изменяется на величину $\hbar\varkappa/m$ ($\varkappa$ — волновой вектор звука, m — эффективная масса электрона). За время между столкновениями τ электрон получает среднее смещение $\Delta = \hbar\varkappa\tau / m$. При столкновении скорость электрона меняется, и он «забывает» о полученном импульсе. Поэтому акустический ток

$$\mathbf{J} = e\hbar\varkappa\tau\nu/m, \tag{1}$$

где ν — число звуковых квантов, поглощаемых электронами в единицу времени, e — заряд электрона.

Учитывая, что $Q_0 = \hbar\omega\nu$ (ω — частота звука), представим величину Q в форме

$$Q = Q_0(1 + \varkappa\mathbf{v}_d / \omega), \tag{2}$$

здесь $\mathbf{v}_d = e\tau\mathbf{E} / m$ — скорость дрейфа.

Таким образом, если векторы $\varkappa$ и $\mathbf{v}_d$ антипараллельны, а скорость дрейфа v_d больше фазовой скорости звука s, то коэффициент поглощения звука $\Gamma = Q / W$ оказывается отрицательным ($W = \rho\omega^2u_0^2\, sV_0 / 2$ — поток энергии в звуковой волне, ρ — плотность кристалла, V_0 — его объем, u_0 — амплитудное значение смещения в звуковой волне). Это является следствием неравновесности распределения электронов в электрическом поле. Величина $\hbar(\omega + \varkappa\mathbf{v}_d)$ представляет собой среднее по распределению изменение кинетической энергии электрона при поглощении звукового кванта. Если эта величина отрицательна, то вероятность испускания кванта становится больше вероятности поглощения и происходит вынужденное черенковское излучение звука.

Наиболее подходящими кристаллами для усиления звука, по-видимому, являются полуметаллы типа висмута. При низких температурах коэффициент решеточного поглощения в висмуте Γ_p относительно мал, а электронный коэффициент $\Gamma_0 = Q_0 / W$ довольно велик. В то же время джоулева мощность $P = nmv_d^2/\tau$ сравнительно невелика, поскольку концентрация электронов n и их эффективная масса m в висмуте малы, а τ — велико.

Можно показать, что выражение (2) для Q остается справедливым и при наличии магнитного поля $\mathbf{H} \perp \mathbf{E}$. При этом в случае $\Omega\tau \gg 1$ ($\Omega = eH / mc$; c — скорость света) скорость дрейфа v_d равна холловской скорости cE / H и направлена поперек полей $\mathbf{E}$ и $\mathbf{H}$, а величина Q_0 должна быть вычислена с учетом магнитного поля. Оказывается, что в случае $\varkappa R \ll 1$, $\varkappa l \gg 1$ ($R = v_F / \Omega$; $l = v_F\tau$; v_F — скорость Ферми) выражение для коэффициента поглощения звука имеет вид

$$\Gamma = \left(1 + \frac{v_d}{s} \cos \varphi\right) \Gamma_0 / |\cos \vartheta|, \tag{3}$$

где φ — угол между векторами $\varkappa$ и $\mathbf{v}_d$; Γ_0 — коэффициент поглощения при $E = H = 0$ [2]; угол ϑ между векторами $\varkappa$ и $\mathbf{H}$ должен удовлетворять условиям

$$|\cos \vartheta| > s / v_F, \qquad |\cos \vartheta| \gg 1 / \varkappa l. \tag{4}$$

Таким образом, коэффициент усиления звука при наличии сильного магнитного поля может быть в отношении v_F/s больше, чем в его отсутствие. Большие по абсолютной величине значения коэффициента Γ в принципе позволяют использовать этот эффект не только для усиления, но для генерирования ультразвука высоких частот.

Следует заметить, что во многих случаях для вычисления коэффициента Γ недостаточно линейного по E приближения. Примерами могут служить случай полупроводника, когда существенно нагревание электронного газа в электрическом поле, и случай магнитного поля, когда поглощение и испускание звука имеет резонансный характер. Однако эти вопросы выходят за рамки настоящей заметки и будут рассмотрены в специальной работе.

Ленинградский физико-технический институт
Академии наук СССР

Поступило в редакцию
17 января 1962 г.

Литература

[1] A. R. Hutson, J. H. Mc Fee, D. L. White. Phys. Rev. Lett., 7, 6, 1961.
[2] А. И. Ахиезер, М. И. Каганов, Г. Я. Любарский. ЖЭТФ, 32, 837, 1957.

Linofilm
Bidirectional Single-Pass Translation

Against/concerning Possibility of Strengthening of Ultrasound In Semi/halfmetals In Electric Field

R. F. Kazarinov, V. G. Skobov

Interaction of/by sound wave with electrons of conductance in crystal brings/lists to dissipation of/by sound energy. With this/besides essential influence on passage of sound can render presence of external electric field E. In work *Khatsona*, Poppy-Fie and *Uayta* [1] detected interesting phenomenon: strengthening of ultrasound in semiconductor CdS in electric field. In this crystal is strong piezoelectric interaction of electrons with sound. However principal possibility of strengthening of sound by conduction electrons in electric field does not depend on concrete character of interaction. Effect has following nature. In absence of electric field electrons absorb sound energy Q_0. This brings/lists to appearance of electronic acoustic current J, proportional Q_0. Therefore sound energy, absorbed electrons in unit of time, in linear along/by E approximation/approach eat/is $Q = Q_0 + JE$. If value E such that $Q < 0$ that takes place strengthening of sound by electrons.

At/during absorption of sound quantum velocity of an electron changes on magnitude $h\chi/m$ (χ–wave vector of sound, m–effective mass of electron). During the time between collisions τ electron receives average displacement $\Delta = h\chi\tau/m$. At/during collision velocity of an electron changes, and he/it "forgets" concerning received pulse. Therefore acoustic current

$$J = eh\chi\tau\nu/m, \qquad (1)$$

where ν–number of sound quanta, absorbed electrons in unit of time, e–charge of electron.

Considering that $Q_0 = h\omega\nu$ (ω–frequency of sound), will present magnitude Q in form

$$Q = Q_0(1+\chi v_d/\omega), \qquad (2)$$

here $v_d = e\tau E/m$–speed of drift.

Thus, if vectors χ and v_d *antiparallelny*, but speed of drift v_d *bolshe* phase speed of sound s, that absorption coefficient sound $\Gamma = Q/W$ appears negative ($W = \rho\,\omega^2 u_0^2 s V_0/2$–flow of energy in sound wave, ρ–density of crystal, V_0–him/his/it/its volume, u_0–peak value of displacement in sound wave). This is result/investigation unequilibriumness of distribution of electrons in electric field. Magnitude $h(\omega + \chi v_d)$ represents average along/by distribution change of/by kinetic energy of electron at/during absorption of sound quantum. If this magnilude negative, that probability of emission of quantum turns *bolshe* probability of absorption and occurs forced Cerenkov radiation of sound.

The most suitable crystals for strengthening of sound, apparently, are semi/halfmetals of type of bismuth. At/during low temperatures coefficient of lattice absorption in bismuth Γ_p relatively small, but electronic coefficient $\Gamma_0 = Q_0/W$ enough great. At the same time joule power $P = nmv_d^2/\tau$ comparatively small, inasmuch as concentration of electrons n and them/their effective mass m in bismuth small, but τ–greatly.

I is possible to show that expression (2) for Q remains just and in the presence of of magnetic field $H \perp E$. With this/besides in case $\Omega\tau >> 1$ ($\Omega = eH/mc$: c–speed of light) speed of drift u_d equal hall*khollovskoy* speed *cE*/*H* and directional/directed across fields E and H, but magnitude Q_0 should be calculated taking into account magnetic field. Turns out that in case $\chi R<<1, \chi l>>1$ ($R=v_F/\Omega$; $l=v_F\tau$; v_F –speed of Fermi) expression for absorption coefficient sound has the form of

$$\Gamma=(1+v_d/s\cos\phi)\,\Gamma_0/|\cos\theta|, \qquad (3)$$

where ϕ–angle between vectors χ and v_d ; Γ_0–absorption coefficient at/during $E=H=0$ [2]; *angle* θ *between vectors* χ *and H should/owe satisfy conditions*

$$|\cos\theta|>s/v_F\ ,\ |\cos\theta|>>1/\chi l \qquad (4)$$

thus, amplification factor sound in the presence of of strong magnetic field can be with respect to v_F /s greater than in its absence. Large along/by absolute value of value of coefficient Γ in principle allow to use this effect not only for strengthening, but for generating of ultrasound of high frequencies.

One should note that in many cases for calculation of coefficient Γ insufficiently linear along/by *E* approximation/approach. Examples can serve case of semiconductor, when considerably heating of electronic gas in electric field, and case of magnetic field, when absorption and emission of sound has resonance character. However these questions appear for/after frame of/by present/real note and will be considered in special work.

Leningrad physicotechnical institute of Academy of sciences of/by USSR

Proceeded in editorial office 17 January 1962 *g*

Literature

[1] A. R. Hutson, J. H. McFee, D. L. White. Phys. Rev. Lett., 7.6. 1961.
[2] a. I. Axiezer, M. I. Kaganov, G +. Lhbarski: j"TF, 32, 837. 1957.

Сообщение ТАСС

В ПОЛЕТЕ «МАРС-1»!

Пролетарии всех стран, соединяйтесь!
Коммунистическая партия Советского Союза
ПРАВДА
Орган Центрального Комитета Коммунистической партии Советского Союза
Газета основана 5 мая 1912 года В. И. ЛЕНИНЫМ | № 306 (16162) | Пятница, 2 ноября 1962 года | Цена 3 коп.

В соответствии с программой исследований космического пространства и планет солнечной системы 1 ноября 1962 года в Советском Союзе осуществлен запуск космической ракеты в сторону планеты Марс.

Подобный запуск осуществлен впервые.

Последняя ступень усовершенствованной ракеты-носителя вывела на промежуточную орбиту тяжелый искусственный спутник Земли, с борта которого была запущена космическая ракета на траекторию движения к планете Марс.

На борту космической ракеты установлена автоматическая межпланетная станция «Марс-1» весом 893,5 кг. Полет автоматической станции до планеты Марс будет продолжаться более семи месяцев.

Основными задачами пуска автоматической станции «Марс-1» являются:

— проведение длительных исследований космического пространства при полете к планете Марс;

— установление межпланетной космической радиосвязи;

— фотографирование планеты Марс с последующей передачей полученных фотографий поверхности Марса на Землю по радиоканалам.

Включение телеметрической, измерительной и научной аппаратуры производится автоматически, в соответствии с программой полета и по радиокомандам с Земли.

Слежение за полетом автоматической станции, определение параметров ее траектории, прием на Земле научной информации осуществляются специальным измерительным комплексом и центром дальней космической радиосвязи.

Предварительные результаты обработки измерительной информации, проведенной в координационно-вычислительном центре, показали, что движение автоматической станции «Марс-1» происходит по траектории, близкой к расчетной. 2 ноября в 10 часов московского времени станция будет находиться на расстоянии 237 тысяч километров от Земли над точкой земной поверхности с координатами 37 градусов западной долготы и 48 градусов северной широты.


Фотомонтаж А. Житомирского.

Вся аппаратура, установленная на борту автоматической станции «Марс-1», работает нормально.

Запуск автоматической межпланетной станции «Марс-1» является дальнейшим этапом в изучении космического пространства и планет солнечной системы.

Tape No. 2154 R/89

Soviet racket/rocket flies to Mars

Communication of/by TASS

In accordance with program of researches of cosmic space and planets of/by solar system 1 November 1962 year in Soviet Union realized starting of space rocket in the direction planet Mars.

Similar starting is realized for the first time.

Last step of/by improved racket/rocket carrier led out on intermediate orbit heavy artificial satellite of Earth, from aboard of that was started space rocket on trajectory of movement to planet Mars.

On board space rocket fixed automatic interplanetary station "Mars-1" weight 893,5 kilogram Flight of/by automatic station to planet Mars will be extend/continued more seven month().

Main problems of starting of/by automatic station "Mars-1" are :

-- effectuation of long researches of cosmic space at/during flight to planet Mars;

-- establishment of/by interplanetary cosmic radio communication;

photographing of planet Mars with/from subsequent transmission received photographs of surface of Mars on Earth by radiocanals.

Switching/inclusion of/by telemetric, measuring and scientific apparatus is produced automatically, in accordance with program of flight and by radioteam/commands from Earth.

Watching for/after flight of/by automatic station, determination of parameters her/it/hers/its trajectory reception on Earth of/by scientific information are realized special measuring complex and center of/by distant cosmic radio communication.

Preliminary results of processing of/by measuring information, passed in koordinatsnonno-calculating center, showed that movement of/by automatic station "Mars-1" occurs along/by trajectory, near to calculation/crew-2 November in 10 hours of Moscow time station will is on distance 237 thousands of kilometers from Earth above point of/by terrestial surface with coordinates 37 degrees of/by West-longitude and 48 degrees of/by north latitude.

All apparatus, fixed on board automatic station "Mars 1", works normally.

Starting of/by automatic interplanetary station "Mars 1" is further/subsequent stage in study of cosmic space and planets of/by solar system.

APPENDIX II. Samples of Multipass Translation

Excerpt from PRAVDA of September 7, 1960 -- Page 5

Если бы Соединенные Штаты и Советский Союз улучшили связи между своими народами, то, вероятно, не было бы такого сильного взаимного антагонизма и могли бы быть созданы условия, чтобы средства, идущие сейчас на военное производство, использовались в мирных целях.

Теперь позвольте нам рассмотреть вопрос о политике воздушной разведки, проводимой правительством США.

Наше первое знакомство с этой политикой произошло в то время, когда мы служили в военно-морских силах США с 1951 по 1954 год. В тот период мы оба являлись техниками по связи на различных станциях радиоперехвата военно-морских сил США.

Если судить по недавним заявлениям правительства США, то оно проводит политику разведывательных полетов вдоль границ и над территорией коммунистических стран якобы лишь в течение последних четырех лет. Однако мы желаем заявить, что подобные полеты проводились также и в период с 1952 по 1954 год, когда мы служили на станции радиоперехвата военно-морских сил США в Камисее близ Иокогамы в Японии.

Перед каждым разведывательным полетом военных самолетов США вдоль китайских и советских дальневосточных границ в Камисею и другие радиоперехватывающие станции посылалась совершенно секретная телеграмма с указанием времени и маршрута такого полета.

В указанное время приемники-мониторы на этих станциях настраивались на частоты, используемые радарными станциями разведуемой страны, в данном случае Советского Союза или коммунистического Китая. Одновременно радиопеленгаторы настраивались на эти же частоты, чтобы определить расположение радарных станций.

Excerpt from PRAVDA of September 7, 1960 -- page 5

If United States and Soviet Union improved connections between own peoples, that, is probable, would not be such strong mutual antagonism and could be created conditions, that means going now for/to military production, are used in peaceful goals.

Now allow us to consider question about politics/policy of air reconnaissance, conducted by government of USA.

Our first acquaintance with this politics/policy happened in that time, when we served in naval forces of USA with/from 1951 to 1954 year. In that period we both were communication technicians on different stations of radio-interception of naval forces of USA.

If one judges by recent declarations of government of USA, that it conducts politics/policy of reconnaissance flights along boundaries and above territory of communist countries supposedly only in course/current of last four years. However we desire to declare, that similar flights were conducted also and in period with/from 1952 along/by 1954 year, when we served on station of radio-interception of naval forces of USA in Kamisei near Yokohama in Japan.

Before each reconnaissance flight of military airplanes of USA along Chinese and Soviet Far-Eastern boundaries in Kamisei and other radio-interception station absolutely secret telegram was sent with indication of time and route of such flight.

In showed time receiver-monitors on these stations were tuned to frequencies, used by radar stations of reconnoitered country, in given case of Soviet Union or communist China. Simultaneously radio direction finders were tuned to the same frequencies, in order to determine location of radar stations.

Russian Text For Multipass Translation

Автоматический перевод языка или "la plume de ma tante"

Гильберт В. Кинг.

Научно-исследовательский центр ИВМа

Все умеют переводить с одного языка на другой, но никто не знает каким образом это осуществляется. Разные подходы могут иллюстрироваться, если переводить с французского на английский выражение

"la plume de ma tante".

Один возможный перевод осуществляется выражением

"the pen of my aunt".

Другой перевод будет выражением

"my aunt's pen".

Можно сказать, что выбор является просто вопросом обычая. Если машинный перевод будет использоваться для чтения, необходимо вырабатывать возможно менее напыщенные конструкции. Собственно говоря, выражение не может переводиться, если контекст не дается, как в примерах, указанных на рисунке "В".

На рисунке "В" нужно идентифицировать выражение

"tante a la mode de Bretagne"

целиком, чтобы переводить правильно слово "tante". Наоборот, если переводчик будет рассматривать каждое слово отдельно, он запутается в двусмысленности. исунки, показывающие записи в словаре для каждого слова, указывают, что обширная двусмысленность существует во всех случаях, кроме слова "ma". Слова вроде предлога "de" имеют многочисленные возможные правильные переводы и для каждого из них потребовалось бы писать целые сочинения.

Слово "ma" очерчивается с достаточной язностью и, на самом деле, оно дает важный предварительный ключ к разгадке грамматического разбора предложения.

Слово "tante", являющееся только существительным, не подвергается двусмысленности. Оно имеет различные значения, но все, кроме слова "aunt" встречаются только в особо специальных контекстах.

Основной проблемой перевода является решение двусмысленностей путем установления контекста. Двусмысленность решается окружающими словами в семидесяти пяти процентах случаев типичного текста, т.е. не написанного лингвистами, чтобы сбить с толку переводчика. Например, слово " ma" в выражении " ma" plume" показывает, что слово "plume" является существительным а не глаголом. Но слово"la", являющееся само по себе двусмысленным словом, не определяет значения слова "plume", потому что контекст мог бы быть выражением "il la plume", "he plucks it -the chicken".

Если ищем слово "plume" в обыкновенном словаре, находим, что оно является существительным женского рода с разными возможными переводами. На рисунке "C", выбор перевода зависит от семантического контекста и такой контекст является, на самом деле, строже в компиляции контекста чем обыкновенный словарь. Однако, замечаем, что третье лицо единственного числа настоящего времени глагола " plumer " пишется " plume", и, фактически, не включается в обыкновенном словаре, где можно найти лишь каноническую форму, т.е. инфинитив " plumer". Однако, нельзя вносить все парадигмы склоняемых форм всех слов, потому что невозможно разграничить разумно ожидаемое от неверсятного или невозможного. Например, окончание " able" может присоединяться к многим глагольным окончаниям, но добавление этого окончания к слову " plume", с известным смыслом, остается под вопросом. Переводчик должен несомненно уметь распознавать приставки и суффиксы и отличать их друг от друга. Основа " plum" может иметь следующие окончания для форм существительного

"e" - единственное число
"es" - множественное число

и тоже глагольное окончание

"e" - третье лицо единственного числа

среди дюжин других глагольных окончаний.

Среди них, окончание "erai" в известной песне "Alouette, " je te plumerai la tete" опознает, что основа является только глаголом, а не существительным, между тем как "e" и " es" не будут решать двусмысленности между существительным и глаголом.

Следовательно, включение в словарь всех лингвистических единиц входного языка, например, основ, окончаний, приставок, суффиксов, полу-идиоматических выражений является первым заданием лексикографа. Заданием лексикографа является рассечение текста в эти, и только в эти лингвистические единицы.

Это осуществляется принципом длиннейшей тождественности. Предложение из вводимого текста находящееся в регистре обозначается на рисунке"E". Показатель находится на точке указывающей степень предварительной разработки текста, т.е., в нашем римере, перед артиклем "la".

Теперь поиск в словаре проводится наугад, но только с точки, находящейся прямо за записями, начинающимися с артикля " la " , и словарь пристально разглядывается в обратном направлении исчерпывающим образом.

Соответствующая страница показывается в весьма упрощенной форме на рисунке "F". Длиннейшая тождественность является здесь простым словом " la " , и сведения об этом слове читаются в регистре, согласно рисунку "C".

Число с заметкой "shift" показывает, что функция этой записи тоже формулирует, что совпадение тождественности достигается на двубуквенном слове, плюс интервал. После этого, показатель подвигается тремя буквами вправо, и интервал между словами считается одним знаком.

Теперь поиск в словаре проводится с точки, находящейся точно за словом "plume" , независимо от последующего слова, скажем, на слове "plupart " . Каждая запись испытывается путем пристального разглядывания в обратном направлении, но полное совпаведие тождественности достигается только для записи " plum ".

Показатель подвигается теперь четырьмя буквами вправо, согласно рисунку "H". Следующий поиск идентифицирует " e en", слово " en " являющееся интервалом. Слово " ma " и слово " tante " рассматриваются подобным же образом. Однако, выражение " ma tante " , имеющее идиоматическое значение идентифицируется как единица, хотя.....

Russian Multipass Translation

Automatic translation of language or "la plume de ma tante"

Gilbert W. King

Scientific research center of IBM

All know how to translate from one language into another, but no one knows in what manner this is realized. Different approaches can be illustrated, if one translates with/from French into English expression

"la plume de ma tante".

One possible translation is realized by expression

"the pen of my aunt".

Another translation will be expression

"my aunt's pen".

It is possible to say, that choice is simply question of usage. If machine translation is used for reading, it is indispensable to work out least possible stilted constructions. Properly speaking, expression cannot be translated, if context is not given, as in examples, showed on figure "B".

On figure "B", it is necessary to identify expressions

"tante a la mode de Bretagne"

entirely, in order to translate correctly word "tante". On the contrary, if translator considers each word separately, he/it will be confused in ambiguities. Figures, indicating entries in dictionary for each word, indicate, that extensive ambiguity exists in all cases, except word "ma". Words like preposition "de" have numerous possible correct translations and for each of them would be required to write complete essay.

Word "ma" is outlined with sufficient clarity and, in fact, it gives important preliminary key to/for solution of parsing of sentence/ proposition.

Word "tante" being only noun, is not subjected to ambiguity. It has different meanings, but all, except word "aunt" are encountered only in particularly special contexts.

Basic problem of translation is solution/decision of ambiguities through establishment of context. Ambiguity is resolved by surrounding words in seventy five per cent of cases of typical text, i.e., not written linguists, in order to stump translator. For instance, word "ma" in expression "ma plume" indicates, that word "plume" is noun and not verb. But word "la" being by itself ambiguous word, does not determine meaning of word "plume", because context could be expression "il la plume", "he plucks it - the chicken".

If we look for word "plume" in ordinary dictionary, we find, that it is noun feminine gender with different possible translations. On figure "C" choice of translation depends on semantic context and such context is, in fact, more severe in compilation of context than ordinary dictionary. However, we notice, that third person/face of singular of present tense of verb "plumer" is written" plume, and, in fact, is not included in ordinary dictionary, where it is possible to find only canonical form, i.e., infinitive "plumer". However, it is impossible to enter all paradigm of declined forms of all words, because it is impossible to demarcate reasonably expected from, improbable or impossible. For instance, ending "able" can be added to/for many verbal endings, but addition of this ending to/for word "plume", with known meaning, remains in doubt. Translator must undoubtedly know how to discern prefixes and suffixes and differentiate them from one another. Stem "plum" can have following endings for forms of noun

"e" -- singular
"es" -- plural

and also verbal ending

"e" -- third person/face of singular

among dozens of other verbal endings.

Among them, ending "erai" in known song "Alouette", "je te plumerai la tete" will recognize, that stem is only verb, and not noun, whereas "e" and "es" will not resolve ambiguity between noun and verb.

Consequently, inclusion in dictionary of all linguistic units of input language, for instance, stems, endings, prefixes, suffixes, semi-idiomatic expressions is first task of lexicographer. Task of lexicographer is dissection of text in these, and only in these linguistic units.

This is realized by principle of longest match. Sentence/ proposition out of introduced text found in register is designated on figure "E". Indicator is found on point indicating degree of preliminary working out of text, i. e., in our example, before article "la".

Now search in dictionary is conducted at random, but only with/from point, found directly after entries, started from article "la", and dictionary is scanned in backward direction exhaustively.

Corresponding page is indicated in quite simplified form on figure "F". Longest match is here simple word "la", and information about this word is read in register, according to figure "C".

Number with notation "shift" indicates that function of this entry also formulates, that match is obtained on two-letter word, plus interval. After this, indicator is moved by three characters to the right, and interval between words is counted one sign.

Now search in dictionary is conducted with/from point, found exactly after word "plume", independently from following word, let us say, on word "plupart". Each entry is tested through scanning in backward direction, but full match is obtained only for entry "plum".

Indicator is moved now by four characters to the right, according to figure "H". Following search identifies "e en", word "en" being interval. Word "ma" and word "tante" are considered in the same manner. However, expression "ma tante" having idiomatic meaning is identified as unit, though ....