R E P O R T R E S U M E S

ED 011 114 AL 000 106

PARSING BY MATRIX--A DEVELOPMENT IN SYNTACTIC ANALYSIS OF RUSSIAN. RESEARCH IN MACHINE TRANSLATION.

BY- STEIGER, AMELIA JANIOTIS
WAYNE STATE UNIV., DETROIT, MICH.

PUB DATE 31 AUG 66

EDRS PRICE MF-$0.18 HC-$2.76 69P.

DESCRIPTORS- *RUSSIAN, *SYNTAX, *MACHINE TRANSLATION, COMPUTATIONAL LINGUISTICS, COMPUTER PROGRAMS, MACHINE TRANSLATION, MT PROJECT, DETROIT, FULCRUM TECHNIQUE, PARSE, HYPERPARSE

RESEARCH IN SYNTACTIC ANALYSIS OF RUSSIAN, WHICH WAS DEVELOPED IN A PROGRAM FOR COMPUTER-AIDED RUSSIAN-ENGLISH TRANSLATION, IS DESCRIBED. THE CORPUS CONSISTED OF 15 RUSSIAN MATHEMATICAL ARTICLES. THE THEORY USED IS THE "FULCRUM" APPROACH OF BUNKER-RAMO, BUT THE COMPUTER IMPLEMENTATION HAS DEVELOPED ALONG DISTINCT LINES. THREE TYPES OF SYNTACTIC ROUTINES ARE DESCRIBED IN THE ORDER OF THEIR APPLICATION--BLOCKING ROUTINES, PROFILING, AND PARSING (PARSE, HYPERPARSE). ALTHOUGH THE IMPROVED PARSING ROUTINE, HYPERPARSE, AND THE AUXILIARY DICTIONARY USED WITH IT ARE A FIRST APPROXIMATION TO SATISFACTORY LANGUAGE TRANSFER, ADDITIONAL CODING IS NEEDED FOR IMPROVED QUALITY OF TRANSLATION. (KL)

ED011114

Research in Machine Translation

PARSING BY MATRIX:

A Development in Syntactic Analysis of Russian

by

Amelia Janiotis Steiger

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Wayne State University
Detroit, Michigan

AL 000 106

AL 000 106

Research in Machine Translation

PARSING BY MATRIX:

A Development in Syntactic Analysis of Russian

by

Amelia Janiotis Steiger
Wayne State University

August 31, 1966

DISTRIBUTION OF DOCUMENT UNLIMITED

This study was supported and sponsored by:

Department of the Navy
Office of Naval Research
Research and Studies Program
Project No. NONR-2562(00)

ACKNOWLEDGEMENT

The author is indebted to Wayne T. Watson, formerly of the programming staff of the Wayne State University machine translation project, for his assistance, encouragement, and interest in the development of this paper.

TABLE OF CONTENTS

PREFACE

There are almost as many approaches to syntactic analysis as there are groups working in the machine translation field. Of these groups, the following have worked on analysis of Russian syntax: Texas, Berkeley, Harvard, RAND, Georgetown, IBM, Bunker-Ramo, and Wayne State University. The approaches of these groups are summarized in the following paragraphs.

Despite the differences in theory and method which exist among the groups, they all have the goal of analyzing the structure of the Russian sentence in order to effect a translational transformation into English. Once the syntactic analysis is accomplished for Russian, semantic features can be brought into the program for the translation stage.

The technical staff of the Linguistic Research Center at Texas consists of three main groups. One of these, viz. Descriptive Linguistics, functions mainly to provide a description of the structure of each language to be used in the translation system. As L.W. Tosh has indicated, "Our approach to describing language structure is a stratificational one."[1] He goes on to point out that this approach comprises three levels of analysis - lexical, syntactic, and semantic - and that the level of greatest interest to the linguist is the syntactic, on which those structural elements which account for number, tense, agreement, and word order are analyzed. In discussing research procedures, he mentions that research on language structures is text-oriented, and although they use general features of Russian structure, they proceed from textual occurrences. They select text that has been translated, using first the Russian alone for monolingual analysis, and then proceeding to the translation in order to construct a transfer grammar.

Sydney M. Lamb in his syntactic research at Berkeley, has been primarily concerned with developing " a system for tactic analysis in general", i.e. a description of arrangements. He states that "... the term syntax is traditionally used with reference to arrangements on the morphemic stratum."[2] He specifies

the form of the syntactic description as follows: "The syntax may be completely described by a list of distribution classes of items with the membership of each, and a list of constructions. A construction is characterized by specification of (1) the distribution classes which enter into it and their relative order, (2) the distribution-class membership of the constitutes."[2] No computer implementation of the method has been announced.

The Harvard group uses the Predictive Syntactic Analysis Technique. Murray Sherry has written: "The method of predictive syntactic analysis is based on the premise that a Russian sentence can be scanned from left to right, and that at any point in this process it is possible both to determine the syntactic structure of the word under consideration on the basis of the predictions made during the analysis of the words to its left, and to predict the syntactic structures which will be encountered to the right of the current word."[3] He points out further: "Predictions of syntactic structures are stored in a prediction pool which behaves somewhat like a pushdown store, a linear array of storage elements in which information is entered or removed from one end only, in accordance with a 'last-in-first-out' principle. New predictions are always entered at the top of the prediction pool, and the predictions are nested starting at the top of the pool and proceeding downward. The topmost prediction in a pool need not necessarily be the next prediction to be fulfilled."[3] It can be seen that at an intermediate point in the analysis of a sentence, the pool contains a set of predictions which are generated by the processing of the preceding words and which are to be fulfilled by the remaining words.

Another well known approach to syntactic analysis is employed by the RAND group and is based on dependency theory as elaborated by David G. Hays. This method, which Hays calls "sentence structure determination", seeks to establish dependency relationships between text occurrences in the sentence. The analysis shows the connections among words in a sentence, where certain

words are said to have other words dependent on them. There are five areas of dependency: subjective, complementary, adjectival, modal, and modificational. Hays has stated: "Dependency theory is actually a characterization theory, not necessarily associable with any empirical method or principle. It is a theory of grammars, with abstract mechanisms for characterizing sets of utterances and for assigning to them certain structural descriptions, which will be called D-trees."[4]

One of the oldest MT research groups, the Georgetown group uses an approach which is called "General Analysis Technique". This method, as Michael Zarechnak has written, seeks to perform the translation operation "... in terms of a machine-programmable analysis and transfer of successively included constituents in the sentence."[5] Their strategy is to perform three levels of analysis on the sentence. On the first level (morphemic), the individual word is analyzed; on the second level (syntagmatic), blocks of adjacent words related in certain ways are constructed; on the third level (syntactic), the subject(s) and predicate(s) of the sentence are located and analyzed. The levels are not self-contained or independent stages; they are segments of the entire technique. A detailed description of the above procedure has been provided by R.R. Macdonald.[6]

The IBM group presently utilizes a sentence-structure-determination routine which "... attempts to parse source-language sentences: to recognize their various constituents and assign them their position within the tree-like structure of the sentence."[7] It is this routine which epitomizes machine translation research, and is the only linguistic area of MT where there is accommodation to hardware. With respect to their multipass translation system, it has been stated that the "... "... search routine with its pass structure attempts to make provision for the recognition of, on the one hand, the constituent structure of sentences, and, on the other hand, the points where sentences are embedded within others...."[7]

Paul L. Garvin has summarized the "fulcrum" approach of the Bunker-Ramo group, saying that the method "... starts out with the minimum unit - the morpheme (minimum unit of gramatical form) in straight linguistic analysis, the typographical word in language data processing - and considers its gradual fusion into units of increasingly higher orders of complexity, called fused units. A sentence is thus visualized, not as a simple succession of linear components, but as a compound chain of fused units of different orders of complexity variously encapsulated into each other. Syntactic analysis, including the automatic analysis which a machine translation syntax routine must perform, then has as its objective the identification of this encapsulation of fused units by a ascertaining their boundaries and functions."[8]

The approach of the Wayne State University group is identical in theory to that of Bunker-Ramo, although the method of computer implementation used by the Wayne group has developed along distinct lines. Certain major aspects of the Wayne method are elaborated in the paper which follows.

The following bibliography comprises the documentation for the aforementioned summaries of research carried out by the indicated machine translation groups.

1. Symposium on the Current Status of Research (Austin, Texas: Linguistic Research Center, University of Texas, October, 1963).

2. Lamb, Sydney M. "On the Mechanization of Syntactic Analysis," Readings in Automatic Language Processing, ed. by David G. Hays (New York: American Elsevier Publishing Co. 1966) pp. 149-158.

3. Sherry, Murray E. "Comprehensive Report on Predictive Syntactic Analysis," NSF-7, Section I (1961).
For additional background, see also:
Oettinger, Anthony G. Automatic Language Translation, (Cambridge, Massachusetts: Harvard University, 1960).

4. Hays, David G. Dependency Theory: A Formalism and Some Observations, (Santa Monica, California: RAND Corp., RM-4087-PR, July, 1964).
For additional information, see also:
Hays, David G. and Ziehe, T. W., Russian Sentence-Structure Determination (Santa Monica, California: RAND Corporation, RM-2538, April, 1960).
Hays, David G. Grouping and Dependency Theories (Santa Monica, California: RAND Corporation, RM 2646, September, 1960).
_____. On the Value of a Dependency Connection (Santa Monica, California: RAND Corporation, RM-2712-AFOSR, January, 1961).

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

6. Macdonald R.R. (ed.) General Report, 1952-1963 (Washington, D.C.: Georgetown University Machine Translation Research Project, June, 1963).

7. Final Report on Computer Set AN/GSQ-16(XW-2), Volume II, The Linguistic Approach (New York: IBM Corporation, Thomas J. Watson Research Center, September, 1963).

8. Garvin, Paul L. An Informal Survey of Modern Linguistics, (American Documentation, Volume 16, No. 4, October, 1965).
See also:
Garvin, Paul L. "Syntactic Retrieval," Proceedings of the National Symposium on Machine Translation, edited by H.P. Edmundson (Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1961).
_____. "Syntax in Machine Translation," Natural Language and the Computer, edited by P.L. Garvin (New York: McGraw-Hill Book Co., 1963).

For a general survey of activities carried out by various MT research groups in the area of syntax, see Summary of the Proceedings of the Conference of Federally Sponsored Machine Translation Groups on MT-Oriented Syntactic Analysis, Machine Translation Research Group, Wayne State University, 1962.

INTRODUCTORY SUMMARY OF RESEARCH IN MACHINE TRANSLATION CONDUCTED AT WAYNE STATE UNIVERSITY

This is a summary of the research which has been carried out at Wayne State University in developing computer programs to perform syntactic analysis on Russian sentences. This research is an integral part of the effort to develop a computer-aided procedure wherein high quality translation from Russian to English is accomplished through the interaction of man and machine.

A corpus of 15 Russian mathematical articles was selected to provide the raw data for experimentation. Each word in the corpus was entered into a dictionary, along with certain of its grammatical properties in coded form, and at least some of its English translations. Each word was put into at least one of a possible nine syntactically based word classes: nominal, predicative, modifier, infinitive, gerund, adverb, preposition, conjunction, declined relative; homographs* were put into two, three, and even four word classes and coded for their properties in each class. The properties for which each word was coded are a function of the word class; the first five mentioned are more densely coded than the last four. The classes where taken from the Ramo-Wooldridge classification scheme.** The class of nominals includes nouns, proper names, and personal pronouns. The class of predicatives comprises ordinary verbs, short form adjectives and participles, and modals. The class of modifiers is made up of adjectives, participles, numerals, and demonstrative pronouns. The class of adverbs contains particles as well as ordinary adverbs. The class of declined relatives is made up of those pronouns which can be used to introduce a relative clause, e.g.: КОТОРЫЙ, ЧЕЙ, КАКОЙ.

---

*A homograph is a word which can be assigned to more than one word class, e.g. НАДО - 'necessary' - (predicative) and - 'over' - (preposition).

**Grammar Code Format and Syntax Flow Charts, an informal collection of material which appeared around 1959.

The corpus and the dictionary were punched onto cards, and were later put on tapes. Programs were written to update the tapes, as well as to select portions of the corpus, look them up in the dictionary, and format the looked-up tapes so that they could serve as input to the automatic syntactic analysis programs which were run on each sentence of the tape.

The syntactic routines used on automatic sentence analysis are of three types. The first type comprises the blocking routines (nominal, prepositional, governing modifier, predicative, and gerund) which group immediate constituents of a sentence into phrases consisting of a fulcrum word and its dependents. The second type comprises the profiling routine which arranges the sentence constituents into columns according to their expected syntactic function(s) in the sentence. The third type (comprising PARSE and HYPERPARSE), using the sentence predicative as fulcrum, determines the actual syntactic roles of many of the sentence constituents (all unnested nominal blocks and certain unnested prepositional phrases) on the basis of the predicative's complementation patterns which are stored in an auxiliary dictionary.

The syntactic routines are continually being revised to include improvements brought to light by observing the output of various runs. There will be a saturation point when improvements in some areas cause greater difficulties in others. HYPERPARSE will have to be extended to include a greater variety of sentence types in its domain of operation, and also to identify the roles of more of the sentence components. Problems of lexical choice for Russian words which have more than one English equivalent will have to be handled, and this will necessitate semantic studies.

The writing of the syntactic routines has been greatly facilitated by a system, now known as GAPS, which enables the language analyst to write in an interpretive language rather than in machine language.

The arrival of the IBM System/360 will necessitate reprogramming of the entire Wayne State University machine translation system, but it is anticipated that the new system will operate much more efficiently both because of technological improvements and because of the incorporation of certain valuable hindsights.

PARSING BY MATRIX*

*A presentation of this research was made at the Fourth Annual Meeting of the Association for Machine Translation and Computational Linguistics, held at the University of California, Los Angeles, California, August 26-27, 1966.

The principal problem treated by the Wayne State University machine translation project since its inception has been Russian syntax. Syntactic resolution of a sentence is an integral part of the process of translating that sentence from the language in which it is given to another language. The purpose of performing syntactic analysis is to discover the structure of any given sentence, where a sentence is defined as a meaningful sequence of words (or idioms), formulas, and punctuation marks, containing at least one verb or verb substitute.

The analysis performed here entails defining various relationships among words and word classes. Routines to seek pertinent items are then programmed, so that instances of these relationships can be recognized in given sentences. In large part, this is accomplished by utilizing the wealth of morphological information about case, number, and gender inherent in Russian forms and displayed in the grammar code of the forms.

The initial relationships are implicitly defined in the various blocking routines. Each blocking routine is brought into operation when an item in a certain word class is discovered. This item is the fulcrum of the block, and the dependents of this fulcrum are recognized and included in the block when they are adjacent to the fulcrum or separated from it by certain permissible items.

The broader relationships on the sentence level are recognized and marked by either of the two parsing routines, both of which utilize the fulcrum approach with the predicative block serving as fulcrum. In these routines, the proximity of the subsidiary sentence items to the fulcrum is not of importance. All of the candidates for each role which can complement the fulcrum are lined up in parallel, reduced logically, and selected in series.

PART I

THE BLOCKING ROUTINES*

*These blocking routines grew out of routines developed on the basis of the fulcrum approach by Paul Garvin of the Bunker-Ramo Corporation, Canoga Park, California, in Grammar Code Format and Syntax Flow Charts, an informal collection of material which appeared around 1959.

Certain elementary relationships, which are difficult to express explicitly, are implicitly defined by the syntactic blocking routines designed in the project. There are five such routines, executed in the following order:

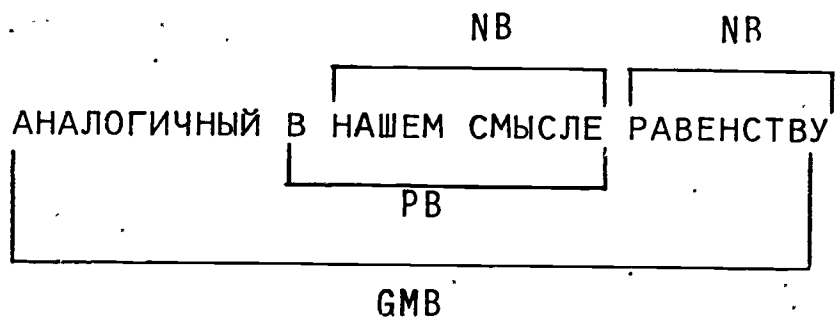
NBR (nominal blocking routine)
PBR (prepositional blocking routine)
GMBR (governing modifier blocking routine)
VBR (predicative blocking routine)
GBR (gerund blocking routine)

Essentially, each routine first seeks the fulcrum element (always an item in the word class or subclass for which the routine is named), and, having found it, attempts to include adjacent items which depend on the fulcrum or on the dependents of the fulcrum, as well as items connecting the dependents.

Prepositional blocks may contain nested nominal blocks as well as nested prepositional blocks. Governing modifier blocks and gerund blocks may contain nested nominal blocks and/or nested prepositional blocks. Nominal blocks and predicative blocks have no nested blocks at present. The concept of nesting may be illustrated in the following example:

In the expression АНАЛОГИЧНЫЙ В НАШЕМ СМЫСЛЕ РАВЕНСТВУ - 'analagous in our sense to the equality' -, the nominal block (NB) НАШЕМ СМЫСЛЕ - 'our sense' - is nested in the prepositional block (PB) В НАШЕМ СМЫСЛЕ - 'in our sense' -, which in turn is nested in the governing modifier block (GMB) АНАЛОГИЧНЫЙ В НАШЕМ СМЫСЛЕ РАВЕНСТВУ.

The structure of this governing modifier block may be illustrated as follows:



The blocking routines, as they are presently formulated, produce blocks composed only of continuous segments of text. The system will eventually have to be expanded, so that phrases which are discontinuous (for example, ПУСТЬ X ВЫПОЛНЯЕТ - 'let X fulfill' - and ПОЗВОЛЯЮТ, В ЧАСТНОСТИ, С ОБЩЕЙ ТОЧКИ ЗРЕНИЯ ОСМЫСЛИТЬ - 'permit, in particular, from a general point of view to interpret' -) can be properly identified.

A description of each of the types of block appears on the following pages.

1) NOMINAL BLOCKING ROUTINE

The nominal blocking routine scans the sentence from left to right until a nominal (noun) is found. The routine then groups the nominal with all of its preceding modifiers (i.e., adjectives, participles, numerals, and certain pronouns), including adverbs modifying the modifiers. The modifiers may be in simple agreement with the nominal; they may be in abnormal agreement, as in the case of numerals greater than one and words like МНОГО - 'much, many' -, НЕСКОЛЬКО - 'several' -; they may be in extended agreement where two or more singular modifiers modify a plural nominal; ПЕРВАЯ И ВТОРАЯ КНИГИ - 'first and second books' -, or one or more plural modifiers modify two or more nominals, the first of which is singular: ХОРОШИЕ КНИГА И КАРАНДАШ - 'good book and pencil' -. In the last case, the second nominal is included in the block.

Those adverbs which are interspersed among a series of modifiers which belong to a nominal are construed to belong to the modifiers and hence to the block. Under certain conditions, adverbs to the left of the leftmost member of such a series of modifiers are included in the block.

Thus, a nominal block is created whose agreement code is that of the nominal, with possible reduction of ambiguity on the basis of the modifiers, or with nominative and/or accusative agreement bits if there is at least one modifier requiring abnormal agreement. The government code of the block is that of the last nominal.

This routine can make the error of linking a governing modifier, which is in the case that it governs, with the following (governed) nominal, to produce a nominal block instead of the correct governing modifier block: ИМЕЮЩИЕ НЕПРЕРЫВНЫЕ ПРОИЗВОДНЫЕ - 'having continuous derivatives' -. When a list of modifiers which must be complemented (e.g. ИМЕЮЩИЙ - 'having' -) is compiled and the dictionary entries are coded, then the nominal blocking

routine could test for this property, and not combine such a modifier with a nominal into a nominal block. This would reduce the frequency of such error.

After creating the nominal blocks, the routine seeks two nominal blocks, having identical agreement codes, which are on opposite sides of a coordinating conjunction. If this is found, the two are combined into one nominal block whose agreement code has the same case(s) as the original ones, but in the plural. For example, ТЕОРЕМА И ФУНКЦИЯ - 'theorem and function' -, where each noun is feminine nominative singular, becomes a block in nominative plural. (In the example: ТАК СКАЗАЛ ГЕЛЬФАНД И ШИЛОВ СОГЛАСИЛСЯ - 'So said Gel'fand and Shilov agreed' -, the combining of ГЕЛФЬАНД И ШИЛОВ - 'Gel'fand and Shilov' -, where both names have identical agreement codes, is incorrect, but it is presumed that such cases are rare.)

There is a question of whether to reduce the requirement that the agreement codes be identical to the requirement that they have non-zero intersection, but this question has not yet been resolved. There is an instance in the corpus (Article V, p. 1, s. 18) where К.И. БАБЕНКО И Г.Е. ШИЛОВЫМ - 'K.I. Babenko and G.E. Shilov' - is not combined, because the agreement code of К.И. БАБЕНКО - 'K.I. Babenko' - includes more cases (since БАБЕНКО - 'Babenko' - is undeclined) than does the agreement code of Г.Е. ШИЛОВЫМ - 'G.E. Shilov' - which is instrumental only. However, if the criterion of combining is reduced to "non-zero intersection of agreement codes", there is the danger that ТЕОРИИ И СИСТЕМЫ - 'theory/theories' and 'system/systems' - would be blocked in the following context: ЭТО УКАЗАНО В ТЕОРИИ (ГЕЛЬФАНДА) И СИСТЕМЫ, ТАКИМ ОБРАЗОМ, ИМЕЮТ СИЛУ., - 'This is proved in the theory (of Gel'fand) and the systems thus hold.' Here, ТЕОРИИ - 'theory' - functions as locative singular, while СИСТЕМЫ - 'systems' - functions as nominative plural; the intersection of agreement codes, however, is genitive singular - nominative plural - accusative plural. In any event, combining the blocks prevents the separate roles of the two blocks from being distinguished.

2) PREPOSITIONAL BLOCKING ROUTINE

The prepositional blocking routine scans the sentence from right to left until a preposition is found. Then, skipping only adverbs and/or nested prepositional blocks, it blocks the preposition with the following a) nominal block, b) (unblocked) modifier, or c) declined relative, provided that the government code of the preposition has positive intersection with the agreement code of a), b), or c). If a declined relative is the object of the preposition, the block is specially marked.

3) GOVERNING MODIFIER BLOCKING ROUTINE

The governing modifier blocking routine scans the sentence from left to right until an unblocked governing modifier is found. The routine then blocks the governing modifier with:

1) governed prepositional blocks

2) governed nominal blocks

3) nominal blocks which agree with the governing modifier.

The following structures are allowed to intervene between the abovementioned three:

a) adverbs

b) any nominal block which can be construed as an adjunct to either the last nominal in the preceding prepositional block or the last nominal in the preceding nominal block, and (potentially) instrumental blocks

c) ungoverned prepositional blocks.

The routine marks nested nominal blocks as to whether they are governed by the governing modifier and/or agree with the governing modifier and/or are adjuncts to a preceding nominal block. It also marks prepositional phrases which are governed.

In the example НЕ УДОВЛЕТВОРЯЕТ ТРЕБОВАНИЯМ, ОБЕСПЕЧИВАЮЩИМ РАЗРЕШИМОСТЬ ИНТЕРПОЛЯЦИОННОЙ ЗАДАЧИ - 'does not satisfy the requirement, ensuring solvability of the interpolation problem' -, since the governing modifier governs the accusative and the dative cases, РАЗРЕШИМОСТЬ - 'solvability' - is marked as governed (since it is nom/acc) and ИНТЕРПОЛЯЦИОННОЙ ЗАДАЧИ - 'interpolation problem' - is marked as an adjunct (since it is gen). In the governing modifier block section of the phrase ПРИ КАКИХ ДОПОЛНИТЕЛЬНЫХ УСЛОВИЯХ, НАЛАГАЕМЫХ НА {λn} - 'under what supplementary conditions, imposed on {λn}' - the prepositional phrase НА {λn} - 'on {λn}' - is marked as a governed prepositional block.

The markings on the nominal blocks constitute a matrix from which it may sometimes be possible to determine whether the governing modifier block is functioning as a nominal block or as a phrase modifying a nominal block which is outside of its boundaries. If a block is marked in position G when it is governed, in position F when it agrees with the governing modifier, and in position A when it is an adjunct, then a nominal block in a governing modifier block may have one of the following vectors:

VECTOR	G	F	A	MEANING
a)	Ø	Ø	Ø	instrumental (where instrumental case is not governed, and governing modifier is not in instrumental case)
b)	Ø	Ø	1	adjunct block
c)	Ø	1	Ø	agreeing (fulcrum) block
d)	1	Ø	Ø	governed block
e)	Ø	1	1	adjunct v agreeing block
f)	1	Ø	1	adjunct v governed block
g)	1	1	Ø	agreeing v governed block
h)	1	1	1	adjunct v agreeing v governed block

The vectors associated with all of the nominal blocks in a governing modifier block (except those which are nested in prepositional phrases) form an n x 3 matrix, where n is the number of nominal blocks in the governing modifier block. If there is exacly one row of type c), the governing modifier block must be made a nominal block, since it contains the agreeing nominal block. The "F" column may then be zeroed out in the rest of the matrix, and the role ambiguity of the remaining nominal blocks may be reduced. If there is exactly one row of type d), the "G" column may then be zeroed out in the rest of the matrix, and a vector of type c) may be sought. (This, of course, holds only when exactly one case is governed.) It will be necessary to investigate additional reduction schemes, so that the governing modifiers blocks which are really nominal blocks, syntactically speaking, may be identified and properly processed.

Modifiers which are marked with the governing bit, but which have no marking for specific governed structures, are allowed. This provision was made because of several examples in text where an ordinary modifier was followed by qualifying phrases, e.g. ПАРАБОЛИЧЕСКИХ В СМЫСЛЕ ПЕТРОВСКОГО СИСТЕМ - 'parabolic in the sense of Petrovsky systems' -. In order to connect ПАРАБОЛИЧЕСКИХ - 'parabolic' - with СИСТЕМ - 'systems' - it is necessary to combine the prepositional phrase В СМЫСЛЕ ПЕТРОВСКОГО - 'in the sense of Petrovsky' -, with the modifier ПАРАБОЛИЧЕСКИХ.

The phrase НЕОТРИЦАТЕЛЬНАЯ НЕПРЕРЫВНАЯ НА СЕГМЕНТЕ [formula] ФУНКЦИЯ - 'negative continuous on the segment [formula] function' -, presents an additional problem, namely that of picking up a sequence of modifiers in a supplemenatry nominal blocking routine to be executed after governing modifier blocking.

One may speculate about the wisdom of subjecting all unblocked modifiers to the governing modifier routine. This, of course would lead to some incorrect results, for example, in the sentence ПРЕДСТАВЛЯЕТСЯ ОЧЕНЬ ЦЕННЫМ В ИССЛЕДОВАНИЯХ ГЕЛЬФАНДА И ШИЛОВА НЕ ТОЛЬКО ВВЕДЕНИЕ ..., - 'Not only the introduction ... is very valuable in the investigations of Gelfand and Shilov' -. Here, the instrumental modifier serves as the complement of the verb rather than a governing modifier, so that if it were classified as a governing modifier, a search for the agreeing nominal would lead to error, since such a nominal does not exist.

4) PREDICATIVE BLOCKING ROUTINE

The predicative blocking routine scans the sentence from left to right until a predicative (finite verb, short form modifier, modal, or special verb form) is found. Then, after searching to the left of the predicative for the negative particle "HE" (skipping adverbs) and including "HE" as the left boundary of the block (if "HE" is found), the routine proceeds to the right of the predicative, identifying and including any temporal auxiliaries and/or infinitive complements found. The agreement code of the block is usually the agreement code of the first predicative, and the government code is that of the last item of the block.

5) GERUND BLOCKING ROUTINE

The gerund blocking routine scans the sentence from right to left until a gerund is found. The routine then blocks the gerund with:

1) governed prepositional blocks
2) governed nominal blocks
3) nominal blocks which agree with the governing modifier.

The following structures are allowed to intervene between the abovementioned three:

a) adverbs
b) any nominal block which can be construed as an adjunct to either the last nominal in the preceding prepositional block or the last nominal in the preceding nominal block, and (potentially) instrumental blocks
c) ungoverned prepositional blocks.

PART II

PROFILE

Once a sentence is blocked, the next step in discovering the structure of the sentence is to analyze it from the standpoint of its fulcrum, the predicative block. In order to do this, it is necessary to put each sentence component, i.e., each block or individual unblocked item, into a list or matrix, according to the potential role(s) of that component in the sentence. This procedure is accomplished by the syntactic routine, PROFILE.

Initially, PROFILE makes the following assignments:

SENTENCE COMPONENT	COLUMN
a) unnested nominal blocks* in the	
i) nominative case	COL I
ii) governed case (under some conditions, see below)	COL II
iii) nominative, genitive, dative, accusative, instrumental cases	COL III
b) predicative blocks	PROFILE
c) governing modifier blocks	DUMP
d) unblocked infinitives	PROFILE
e) gerund blocks	DUMP
f) unblocked adverbs	DUMP
g) unnested prepositional blocks	
i) with declined relative object	PROFILE
ii) other	PREP
h) unblocked conjunctions (including most punctuation)	PROFILE
i) expressions in parentheses	DUMP
j) declined relatives	PROFILE

NOTE: An unblocked modifier presently leads to an error condition.

*Since a nominal block may be in more than one case, it may be entered in more than one column. For example, the noun НОЧИ - 'night/nights' - may be genitive, dative or locative singular or nominative or accusative plural, and may therefore be entered both in COL I and COL III.

It can be seen that the components which are entered into PROFILE are, for the most part, clause determiners. Those which are entered into DUMP are not part of the essential structure of the sentence. Unnested nominal blocks, which are in COL I and COL III, can have the following roles in a sentence:

i) subject

ii) governed predicative complement

iii) adjunct, where adjunct is presently taken to mean:

a) a dependent of a preceding nominal block

b) an adverbial expression in the instrumental case, e.g. ТАКИМ ОБРАЗОМ - 'thus' -, or the accusative (of time) case, e.g. ВСЮ НОЧЬ - 'all night' -.

iv) appositive (This role will be ignored in what follows.)

The prepositional blocks, which are in PREP, can have the following roles in a sentence:

i) governed predicative complement

ii) adverbial expression

iii) adjunct to a preceding nominal block (This role will be ignored in what follows.)

After creating the columns, the routine proceeds to test the column entitled PROFILE in order to ascertain the number of predicative blocks in the sentence. At present, further analysis is done only on those sentences having at most one predicative block. (Sentences having no predicative block are treated as if they had a verb governing the nominative case only.) If the sentence has exactly one predicative block, the block is tested for its government properties. If the block governs at most one case and nothing else, then the routine creates an additional column, COL II, consisting of all unnested nominal blocks in the governed case, and calls on the syntactic routine PARSE for further analysis. If the block has other government properties, then the syntactic routine HYPERPARSE is called upon for further analysis. PARSE and HYPERPARSE are described in the following pages.

PART III

PARSE

If we consider only simple sentences having a predicative block which governs at most one case and no preposition (in which case unnested prepositional blocks will be regarded as adverbial), it is possible to determine all possible interpretations of a sentence by taking into account all possible roles of each unnested nominal block in that sentence. When each nominal block has been assigned one of the three roles: subject, governed predicative complement, or adjunct, then the sentence is said to be 'parsed', and the assignment is called a parsing.

The technique used on this project for finding all possible parsings of a simple sentence having a predicative block with restricted government is to create, modify, and analyze an n x 3 matrix, where n is the number of unnested nominal blocks in the sentence, and 3 is the number of roles which can presently be assigned to a nominal block.

1) The creation of the matrix is accomplished by the syntactic routine PROFILE. COL I contains the potential subjects, i.e., nominal blocks in the nominative case. COL II contains the potential governed predicative complements, i.e., nominal blocks in the governed case. COL III contains the potential adjuncts, i.e., nominal blocks in the cases other than locative. A given nominal block can be entered in as many columns as its agreement code will allow.

Once the matrix is created by PROFILE, the modification and analysis of the matrix are accomplished by the syntactic routine PARSE.

2) The modification is carried out in two stages:

I. Certain grammatical considerations sometimes make it possible to remove entries from COL I and COL III.

A) Col I can be reduced by removing all nominal blocks which do not have the correct number, person, or gender to be the subject of the predicative block of the sentence. Although, НОЧИ - 'night/nights' - has nominative plural bits, it cannot be the subject in a sentence where ДАЕТ - 'gives' - is the predicate, since ДАЕТ is singular. (possible cause of error: compound subject, e.g. ИВАН И КАТЯ ИДУТ - 'John and Kathy are going' -)

B) COL III can be reduced by removing every potential adjunct which is not in the instrumental case (or accusative of time) or is not immediately preceded by a nominal block which governs it. (possible cause of error: governor of adjunct does not immediately precede adjunct, e.g., ПРОБЛЕМА КОШИ ЕДИНСТВЕННОСТИ - 'Cauchy's problem of uniqueness' -)

II. After COL I and COL III are reduced through grammatical considerations, it is possible to reduce COL I and COL II using logical considerations, provided that the following condition is made: In any given parsing, at most one element (nominal block) may be assigned the role of subject, and at most one element may be assigned the role of governed predicative complement. (This means that sentences having compound subjects and/or compound governed predicative complements, e.g., ТЕОРЕМА ГЕЛЬФАНДА И ТЕОРЕМЫ ШИЛОВА ПРЕДСТАВЛЯЮТСЯ - 'The theorem of Gel'fand and the theorems of Shilov are presented' -, cannot be properly parsed. Sentences with compound subjects and/or compound predicative complements can be parsed when the components of the compound block are combined into one block by the NBR, which does so under special circumstances.)

The logical reduction takes place as follows: If there is a nominal block, N, in COL I (or COL II), such that N is not entered in any other column, then all other entries in COL I (or COL II) must be erased.

The reason is:

a) N must be assigned a role

b) N can only be assigned the role associated with COL I (or COL II)

c) the role associated with COL I (or COL II) can be assigned to at most one nominal block

hence N is the only nominal block which can be assigned the role associated with COL I (or COL II), and, in fact, N must always be assigned that role.

NOTE: If there are two such Ns in a column, an error message is written.

In any reduction step, it is imperative to avoid obliterating any nominal block, i.e. erasing it from some column when it is not entered in any other column, for a parsing must assign a role to each unnested nominal block. In PARSE, the obliteration of a nominal block in either stage of the modification portion of the routine causes an error message to be written, and the parsing to be discontinued.

3. The analysis portion of the PARSE routine consists of finding all possible parsings, i.e., all possible ways of choosing at most one nominal block from COL I and at most one nominal block from COL II and as many nominal blocks as necessary from COL III so that exactly one assignment is made for each row in the matrix.

Consider the following matrix:

	COL I	COL II	COL III
A) first unnested nominal block	X		X
B) second unnested nominal block		X	X
C) third unnested nominal block	X	X	X

The set of parsings, where each matrix element is represented by its row (A, B, C) and column (I, II, III), is:

PARSING	SUBJECT	OBJECT	ADJUNCT(S)
1	A,I	B,III	C,III
2	A,I	C,II	B,III
3	C,I	B,II	A,III
4	A,I	NONE	B,III; C,III
5	C,I	NONE	A,III; B,III
6	NONE	B,II	A,III; C,III
7	NONE	C,II	A,III; B,III
8	NONE	NONE	A,III; B,III; C,III

NOTE: The above list of parsings illustrates a feature of the PARSE routine: namely, that the absence of the subject and/or of the governed item is allowed, as long as the parsing utilizes all unnested nominal blocks in the sentence.

The parsing scheme just described was satisfactory for sentences containing predicatives with a limited simple type of government. Before proceeding to the description of the more general parsing routine, HYPERPARSE, it is valuable to consider the more complex types of predicative government.

PART IV

PREDICATIVE GOVERNMENT*

*For an excellent description of complementation, see Andrew S. Kozak, Complementation in Russian: Theory and Application, Memorandum RM-4582-PR (The RAND Corporation, Santa Monica, California), September, 1965.

Predicative government is a term which can roughly be defined as the property possessed by a predicative of requiring complementation by one or more of various structures. Complementation is distinct from modification, and the distinction lies in whether a structure which is a candidate for complementation can be used (almost) universally (in which case it is a modifying structure) or is peculiar either to the predicative being studied or to a proper subset of all predicatives.

The types of structure which can serve as complements to a predicative are:

1) nominal blocks (or modifiers) in*
   - (N) nominative case
   - (G) genitive case
   - (D) dative case
   - (A) accusative case
   - (I) instrumental case
2) prepositional blocks
   (the set of prepositions which are actually in complementary structures has not yet been defined)
3) ЧТО and ЧТОБЫ phrases
4) infinitives
5) КАК phrases
6) adverbs

Infinitive government is handled in the VBR when the infinitive is contiguous to the predicative (allowing intervening adverbs). Types 5) and 6) have not yet been studied extensively.

Let us consider only case and prepositional complementation. ДЕЛАТЬ - 'to do' - can govern accusative (A), dative (D), instrumental (I), and ИЗ - 'out of' + genitive (G). This alone does not specify which combinations of these governed items can occur. It happens that the accusative can occur alone, with the instrumental or with the dative, or with ИЗ + genitive. Also, the accusative can occur with the dative and ИЗ + genitive. This information gives five valid complementation patterns which are represented and exemplified as follows:

*(L) locative case is not a predicative complementation case.

1) A	ОН ДЕЛАЕТ ГОРШОК. – 'He is making a pot.'
2) D and A	ОН НАМ ДЕЛАЕТ НОВОЕ ПРЕДЛОЖЕНИЕ. – 'He is making us a new offer.'
3) A and I	ЭТО ДЕЛАЕТ ГЛИНУ ОГНЕУПОРНОЙ. – 'This makes the clay fireproof.'
4) A and ИЗ + G	ОН ДЕЛАЕТ ИЗ ГЛИНЫ ГОРШОК. – 'He is making a pot out of clay.'
5) D and A and ИЗ + G	ОН НАМ ДЕЛАЕТ ИЗ ГЛИНЫ ГОРШОК. – 'He is making us a pot out of clay.'

(Note that a total of 16 patterns: $\binom{4}{4}+\binom{4}{3}+\binom{4}{2}+\binom{4}{1}+\binom{4}{0}=16$; can be obtained from all combinations of the four governed structures; the enumeration of the valid ones reduces this number to five.)

Predicative government in the original MT dictionary of this project was coded in such a way that all complements for a given predicative are indicated in summary only, without any indication of which combinations of these governed structures can actually occur. When considering predicative blocks with extremely simple government codes (i.e., at most one case governed, and no prepositions governed), the predicative government coding of the original MT dictionary is usable. However, in order to parse sentences having a predicative block with more complex complementation patterns, more comprehensive input information about these patterns is needed.

Predicative government input information now consists of the enumeration of each combination in which the governed structures (nominal blocks in a certain case prepositional blocks where the preposition and the case it governs are specified, and, for later use, clauses introduced by ЧТО - 'that' - and ЧТОБЫ - 'that' - and phrases introduced by КАК - 'as' -) can occur. Each possible combination (including information about whether a subject can occur) is called a pattern, and the totality of patterns for a given predicative is called the pattern set for that predicative. The pattern set for each predicative in the corpus is stored in an auxiliary dictionary.

Following are the coding instructions used in the creation of the auxiliary dictionary. The instructions are followed by sample coding forms for the patterns of ДЕЛАТЬ - 'to do' These forms are kept in a permanent reference file for language-example documentation of each pattern.

Coding Instructions for Verbal Government Patterns

FIRST (CANONICAL FORM) CARD:

cols 1- 6 ID NUMBER

Write the four-digit identification number assigned to the canonical form. The fifth and sixth digits, entitled 'XTRA', are used only for inserting a form. Otherwise they are blank.

cols 7-12 blanks

cols 13-42 RUSSIAN CANONICAL FORM

Write the infinitive form of the verb, or the neuter form of the short form modifier for which the patterns apply. In case a certain form of the set represented by the infinitive or the neuter short form has a special set of patterns, write that form and code its patterns separately.

cols 43-80 blanks

NEXT (PATTERN) CARDS:

cols 1- 6 ID NUMBER

Write the four-digit identification number assigned to the canonical form. The fifth and sixth digits, entitled 'XTRA', are used only for inserting a form. Otherwise they are blank.

cols 7 P

cols 8-10 PATTERN #

Write the two digit number assigned to the pattern being coded. The third digit, entitled 'XTRA', is used only for inserting a pattern. Otherwise it is blank.

cols 11-12 blanks

cols 13 SUBJECT

If the third person singular or the neuter form associated with the canonical form must be impersonal in the pattern being coded, code 'N' in the square alloted; otherwise code 'S'.

cols 14-17 CASE GOVERNMENT

For each case governed by the verb in the pattern being coded, write the appropriate one-digit number, beginning in the first square of the field. Unused squares are to be left blank. See table entitled 'CASE GOVERNMENT CODES'.

cols 18-25 PREPOSITION GOVERNMENT

For each preposition + case pair governed by the verb in the pattern being coded, write the appropriate two-digit number, beginning in the first two squares of the field. Unused squares are to be left blank. See table 'PREPOSITION + CASE GOVERNMENT CODES'.

cols 26,27 ЧТО/ЧТОБЫ CLAUSE GOVERNMENT

If a ЧТО or ЧТОБЫ clause is governed in the pattern being coded, write a 1 in the appropriate square.

cols 28 КАК PHRASE GOVERNMENT

If a КАК phrase is governed in the pattern being coded, write a 1 in the square.

cols 29 INFINITIVE GOVERNMENT

If the verb governs an infinitive in the pattern being coded, then code a 1 in the square.

cols 30-80 blanks

Write a Russian example with an English translation to illustrate the pattern coded above.

CASE GOVERNMENT CODES

NO	CASE
1	nominative case
2	genitive case
3	dative case
4	accusative case
5	instrumental case

AUXILIARY DICTIONARY PREPOSITION AND CASE GOVERNMENT CODES

NO	PREP	CASE
01	В, ВО	+ acc.
02	В, ВО	+ loc.
03	НА	+ acc.
04	НА	+ loc.
05	О, ОБ, ОБО	+ acc.
06	О, ОБ, ОБО	+ loc.
07	ЗА	+ acc.
08	ЗА	+ instr.
09	ПОД, ПОДО	+ acc.
10	ПОД, ПОДО	+ inst.
11	ПО	+ acc.
12	ПО	+ dat.
13	ПО	+ loc.
14	С, СО	+ acc.
15	С, СО	+ gen.
16	С, СО	+ instr.
17	ДО	+ gen.
18	ИЗ	+ gen.
19	ОТ, ОТО	+ gen.
20	ПРО	+ acc.
21	К, КО	+ dat.
22	НАД, НАДО	+ instr.
23	ПРИ	+ loc.
24	ЧЕРЕЗ	+ acc.
25	МЕЖДУ	+ gen.
26	МЕЖДУ	+ instr.
27	ИЗ-ЗА	+ gen.
28	ИЗ-ПОД, ИЗ-ПОДО	+ gen.
29	ДЛЯ	+ gen.
30	У	+ gen.
31	ПЕРЕД, ПЕРЕДО	+ instr.
32	СРЕДИ	+ gen.
33	ПРОТИВ	+ gen.
34	ПУТЕМ	+ gen.
35	ВМЕСТО	+ gen.
36	МИМО	+ gen.

col

1	2	3	4	5	6
ID NUMBER				XTRA	XTRA

RUSSIAN (CANONICAL) FORM (cols 13-42, first card)

col

7	8	9	10	13	14 15 16 17	18 19 20 21	22 23 24 25	26	27	28	29
P											
	PAT-TERN #		XTRA	SUBJ	CASE GOVERNMENT	PREPOSITIONAL GOVERNMENT		ЧТО	ЧТО-БЫ	КАК	INF

EXAMPLE:

RUSSIAN _____

_____

_____

_____

ENGLISH _____

_____

_____

_____

col

1	2	3	4	5	6
ID NUMBER				XTRA	XTRA

ДЕЛАТЬ

RUSSIAN (CANONICAL) FORM (cols 13-42, first card)

col

7	8	9	10	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
P	Ø	1		S	4															
PAT-TERN #			XTRA	SUBJ	CASE GOVERNMENT				PREPOSITIONAL GOVERNMENT								ЧТО	ЧТО-БЫ	КАК	INF

EXAMPLE:

RUSSIAN ОН ДЕЛАЕТ ГОРШОК.

ENGLISH He is making a pot.

col

1	2	3	4	5	6
ID NUMBER				XTRA	XTRA

ДЕЛАТЬ

RUSSIAN (CANONICAL) FORM (cols 13-42, first card)

col

7	8	9	10	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
P	Ø	2		S	3	4														
	PAT-TERN #		XTRA	SUBJ	CASE GOVERNMENT				PREPOSITIONAL GOVERNMENT								ЧТО	ЧТО-БЫ	КАК	INF

EXAMPLE:

RUSSIAN ОН НАМ ДЕЛАЕТ НОВОЕ ПРЕДЛОЖЕНИЕ.

ENGLISH He is making us a new offer.

col

1	2	3	4	5	6
ID NUMBER				XTRA	XTRA

ДЕЛАТЬ

RUSSIAN (CANONICAL) FORM (cols 13-42, first card)

col

7	8	9	10	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
P	Ø	3		S	4	5														
	PAT-TERN #		XTRA	SUBJ	CASE GOVERNMENT				PREPOSITIONAL GOVERNMENT								ЧТО	ЧТО-БЫ	КАК	INF

EXAMPLE:

RUSSIAN ЭТО ДЕЛАЕТ ГЛИНУ ОГНЕУПОРНОЙ.

ENGLISH This makes the clay fireproof.

col

1	2	3	4	5	6
• ID NUMBER				XTRA	XTRA

ДЕЛАТЬ

RUSSIAN (CANONICAL) FORM (cols 13-42, first card)

col

7	8	9	10	13	14 15 16 17	18 19 20 21 22 23 24 25	26	27	28	29
P	Ø	4		S	4	1 8				
	PAT-TERN #		XTRA	SUBJ	CASE GOVERNMENT	PREPOSITIONAL GOVERNMENT	ЧТО	ЧТОБЫ	КАК	INF

EXAMPLE:

RUSSIAN ОН ДЕЛАЕТ ИЗ ГЛИНЫ ГОРШОК.

ENGLISH He is making a pot out of clay.

col

1	2	3	4	5	6
ID NUMBER				XTRA	XTRA

ДЕЛАТЬ

RUSSIAN (CANONICAL) FORM (cols 13-42, first card)

col

7	8	9	10	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
P	Ø	5		S	3	4			1	8										
	PAT-TERN #		XTRA	SUBJ	CASE GOVERNMENT				PREPOSITIONAL GOVERNMENT								ЧТО	ЧТО-БЫ	КАК	INF

EXAMPLE:

RUSSIAN ОН НАМ ДЕЛАЕТ ИЗ ГЛИНЫ ГОРШОК.

ENGLISH He is making us a pot out of clay.

PART V

HYPERPARSE

In the routine called HYPERPARSE, the set of government patterns for the predicative block is looked up in the auxiliary dictionary. Each pattern in the set is compared with the nominal and preposition blocks in the sentence in an attempt to find all possible realizations of the pattern in that sentence. As in PARSE, sentences containing uncombined compounds of nominal blocks (and sentences containing nouns in apposition) will either be parsed incorrectly or else will not be parsed for the reason given in an appropriate error message printed out with the sentence.

This section contains a description of how a blocked, profiled sentence is treated by HYPERPARSE. The sentence ЭТА ТЕОРЕМА ДВОЙСТВЕННОСТИ ПОЗВОЛЯЕТ ДОКАЗАТЬ ЛЕММЫ ТОЛЬКО В ЭТОМ СЛУЧАЕ. - 'This theorem of duality allows (one) to prove the lemmas only in this case.' - is used to illustrate the process.

The blocks in the sentence are:

NOMINAL BLOCKS:

# of NB	cases of singular/plural	word # of LB*	word # of RB**	
1	N/Ø	1	2	ЭТА ТЕОРЕМА
2	G,D,L/N,A	3	3	ДВОЙСТВЕННОСТИ
3	G/N,A	6	6	ЛЕММЫ
4	L/Ø	9	10	ЭТОМ СЛУЧАЕ

PREPOSITIONAL BLOCKS:

word # of LB*	word # of RB**	
8	10	В ЭТОМ СЛУЧАЕ

GOVERNING MODIFIER BLOCKS: none

PREDICATIVE BLOCKS:

word # of LB*	word # of RB**	
4	5	ПОЗВОЛЯЕТ ДОКАЗАТЬ

GERUND BLOCKS: none

*LB = left boundary
**RB = right boundary

The profile of the sentence, as composed by the PROFILE routine, is represented in the printout as follows:

WORD #	RUSSIAN/ENGLISH	PROFILE	COL I	COL II	COL III	PREP	DUMP
1	ЭТА THIS		B		B		
2	ТЕОРЕМА THEOREM		E		E		
3	ДВОЙСТВЕННОСТИ DUALITY/DUALITIES		X		X		
4	ПОЗВОЛЯЕТ ALLOWS+DOES ALLOW	B					
5	ДОКАЗАТЬ TO PROVE	E					
6	ЛЕММЫ LEMMA/LEMMAS		X		X		
7	ТОЛЬКО ONLY						*
8	В IN/INTO					B	
9	ЭТОМ THIS						
10	СЛУЧАЕ CASE					E	
11	. .	*					

NOTE: In the translation field, entries may be separated by a slash (/), an asterisk (*), or a plus sign (+). A slash is used to separate different meanings, an asterisk is used to separate the singular and plural forms of a given meaning, and a plus sign is used to separate two different forms of a verb, e.g. goes + does go. The progressive form ('is going') is not stored in the translations.

In the columns, B indicates "beginning of block," E indicates "end of block", X indicates a one word block, and * indicates an unblocked item.

Internally, PROFILE constructs the following matrix for the unnested nominal blocks in the sentence, where the entries are the word numbers of the left and right boundaries of a block:

NB #	COL I	COL II	COL III
1	1-2		1-2
2	3-3		3-3
3	6-6		6-6
4			

NOTE: COL II is empty because the predicative block (ПОЗВОЛЯЕТ ДОКАЗАТЬ) governs more than one case.

The predicative block in the sentence governs the accusative and dative cases in the following combinations only:

Pattern 1: A

Pattern 2: DA

Henceforth, COL II will be considered to have as many sub-columns as there are governed cases in the pattern under consideration. Hence, we have COL II_A for Pattern 1, and COL II_A and COL II_D for Pattern 2.

Replacing the word numbers of the boundaries by the symbol X to indicate that a nominal block has been entered in a certain position of the matrix, and substituting the symbol S_i for the ith nominal block, and considering only the unnested nominal blocks, we have the following two nominal block matrices for the sentence and the two patterns:

MATRIX FOR PATTERN 1	COL I	COL II_A	COL III
S_1	X		X
S_2	X	X	X
S_3	X	X	X

MATRIX FOR PATTERN 1	COL I	COL II_D	COL II_A	COL III
S_1	X			X
S_2	X	X	X	X
S_3	X		X	X

where S_i is the ith unnested nominal block.

Since S_2 and S_3 do not agree with the predicative block (for the nominative case they are plural while the predicative block is singular), they can be removed from COL I. Also, neither S_1 nor S_3 is preceded by a governing nominal block, nor is either in the instrumental case, and therefore each can be removed from COL III. None of the removals causes obliteration of the S_i from the matrix.

The matrices after the initial modification are now as follows:

MATRIX FOR PATTERN 1	COL I	COL II_A	COL III
S_1	X		
S_2		X	X
S_3		X	

MATRIX FOR PATTERN 1	COL I	COL II_D	COL II_A	COL III
S_1	X			
S_2		X	X	X
S_3			X	

Since S_3 is only in COL II_A, S_2 must be removed from that column. This can be done without obliteration, so the matrices are now as follows:

MATRIX FOR PATTERN 1	COL I	COL II_A	COL III
S_1	X		
S_2			X
S_3		X	

MATRIX FOR PATTERN 1	COL I	COL II_D	COL II_A	COL III
S_1	X			
S_2		X		X
S_3			X	

The matrix for Pattern 1 is finished, and yields exactly one parsing where

S_1 is the subject

S_2 is the adjunct

S_3 is governed by the predicative block

This is the only "correct" parsing, but examination of the matrix for Pattern 2 shows that another parsing is formally, though not semantically possible.

In HYPERPARSE, each sub-column of COL II must have at least one entry, or the pattern cannot apply. Therefore, for Pattern 2 an additional logical operation which leads to the removal of S_2 from COL III is used. The operation can be described by saying that if S_i is the only entry in a given sub-column of COL II, then the entries of S_i are erased elsewhere in the row.

The final matrix for Pattern 2,

PATTERN 2	COL I	COL II_D	COL II_A	COL III
S_1	X			
S_2		X		
S_3			X	

yields exactly one parsing where

S_1 is the subject
S_2 is governed by the predicative block
S_3 is governed by the predicative block

The translation of the sentence when it is parsed in this way, is: 'This theorem allows (one) to prove the lemmas to the duality only in this case.' (The English word order had to be adjusted to allow 'duality' to function as the indirect object of 'prove'.) It should be noted that any indirect object in this context probably has to be animate, and that sometime in the future, this restriction should be coded into the pattern. This will eliminate the semantically unsound parsing here and perhaps in many other cases.

Just as HYPERPARSE creates a matrix for the unnested nominal blocks which are in the cases specified by the pattern being considered, it also creates a matrix of unnested governed prepositional blocks. The PREP column is the source of an NPREP matrix which for each pattern has as many sub-columns as there are governed prepositions in the pattern. Each sub-column must have at least one prepositional block entry corresponding to a governed preposition in an input pattern, in order for the pattern to apply.

The governed prepositions P_1, P_2, and P_3 in a given pattern, may lead to the following matrix in a hypothetical sentence, where each entry comprises the boundaries of a prepositional block:

P_1	P_2	P_3
1- 3	17-18	6-9
12-14		

This means that for one of the governed prepositions, two realizations have been found. The parsings will then be:

(1- 3)(17-18)(6-9)
(12-14)(17-18)(6-9).

When both cases and prepositions are governed in a given pattern, each case parsing must be joined with each preposition parsing to produce the complete set of parsings for the pattern.

The following paragraphs give a verbal summary of the flow charts* of HYPERPARSE:

A. Create a subject column, SUBINT (corresponding to the reduced COL I of PARSE), containing only those unnested nominal blocks which agree with the predicative block through the agreement code and the person bits.
   NOTE: If the pattern being tried allows a subject, SUBINT becomes the subject column (SUBCOL).

B. Create two adjunct columns, NORADJ and SPCADJ (each corresponding to the reduced COL III of PARSE), containing only those nominal blocks which can actually serve as adjuncts in the given sentence. NORADJ contains each unnested nominal block which is (immediately) preceded by a nominal block governing it, or which is in the instrumental case; SPCADJ, of which NORADJ is a subset, contains, in addition, any nominal block in the accusative case.
   NOTE: If the accusative case is governed in the pattern being tried, NORADJ is used as the adjunct column (ADJCOL); otherwise SPCADJ, is used.

C. After a pattern is read in, create a prepositional column $NPREP_p$, p=1(1)P, $P \leq 4$, for each preposition governed in the pattern. $NPREP_p$ contains all unnested prepositional phrases in the sentence such that the preposition is a governed one and the case of its governed nominal block is the one specified for that preposition.

D. Create an object column, $OBJCOL_c$, c=1(1)C, $C \leq 4$, for each case in the pattern. $OBJCOL_c$ contains all unnested nominal blocks in the sentences which are in a given governed case.

E. Check the validity of the matrix created in A, B, and D, i.e. ascertain that every unnested nominal block in the sentence has been entered in at least one of SUBCOL, $OBJCOL_{(1,\ldots,C)}$, or ADJCOL.

*The flow charts are available to persons having special interest in the programming details of this procedure.

F. Reduce the matrix according to the following two schemes:

   a) Check each column for an entry which is not in any other column. If exactly one such entry is found in a column, erase the other entries in the column.

   b) Check each $OBJCOL_{(1,...,C)}$ for having exactly one entry. If this condition exists then erase the other entries in the same row.

G. Find all possible ways of choosing an entry from each of SUBCOL and $OBJCOL_{(1,...,C)}$ such that no two entries chosen are in the same row.
(NOTE: In some cases, the above choices are made without using SUBCOL, e.g., when the predicative in the sentence must be impersonal, or as an alternative when the matrix is valid without SUBCOL, even when a subject is allowed.)
Then ascertain that all rows having entries but which have not been selected, do have an entry in the adjunct column, and select those entries.
Combine the selected entry pattern in G, with the selection of the first row (if it is non-empty) of the prepositional block matrix, and write out the parsing.

Paragraphs A-G are merely a simplified summary of how a successful parsing involving governed cases and prepositional blocks is made. Alternatives used when certain conditions (assumed or ignored in the above) are not met, and the corresponding error messages are shown in detailed flow charts.

The auxiliary dictionary and the improved parsing routine, HYPERPARSE, are only a first approximation to a good language transfer system. The quality will be improved when the preword(s) associated with each governed case and the translation(s) associated with each governed preposition are coded for each government pattern, along with the translation of the predicative for that given context. This coding will reduce the number of meanings to be printed out.

DISTRIBUTION LIST

ONR BASIC RECIPIENTS MAILING LIST

TECHNICAL LIBRARY
DIRECTOR DEFENSE RES. + ENG
ROOM 3C-128, THE PENTAGON
WASHINGTON, D. C. 20301

DEFENSE DOCUMENTATION CENTER — 20
CAMERON STATION
ALEXANDRIA VIRGINIA 22314

CHIEF OF NAVAL RESEARCH — 02
DEPARTMENT OF THE NAVY
WASHINGTON 25, D. C.
ATTN CODE 437, INFORMATION SYSTEMS BRANCH

DIRECTOR, NAVAL RESEARCH LABORATORY — 06
TECHNICAL INFORMATION OFFICER /CODE 2000/
WASHINGTON 25, D. C.

COMMANDING OFFICER, OFFICE OF NAVAL RESEARCH — 10
NAVY =100, FLEET POST OFFICE BOX 39
NEW YORK, NEW YORK, 09599

COMMANDING OFFICER
ONR BRANCH OFFICE
207 WEST 24TH STREET
NEW YORK 11, NEW YORK

OFFICE OF NAVAL RESEARCH BRANCH OFFICE
495 SUMMER STREET
BOSTON, MASSACHUSETTS 02110

NAVAL ORDNANCE LABORATORY
WHITE OAKS
SILVER SPRING 19, MARYLAND
ATTN TECHNICAL LIBRARY

DAVID TAYLOR MODEL BASIN
CODE 042 TECHNICAL LIBRARY
WASHINGTON D. C. 20007

ERIC

NAVAL ELECTRONICS LABORATORY
SAN DIEGO 52, CALIFORNIA
ATTN TECHNICAL LIBRARY

DR. DANIEL ALPERT, DIRECTOR
COORDINATED SCIENCE LABORATORY
UNIVERSITY OF ILLINOIS
URBANA, ILLINOIS

AIR FORCE CAMBRIDGE RESEARCH LABS
LAURENCE C. HANSCOM FIELD
BEDFORD, MASSACHUSETTS
ATTN RESEARCH LIBRARY, CRMXL-R

U. S. NAVAL WEAPONS LABORATORY
DAHLGREN, VIRGINIA 22448
ATTN G. H. GLEISSNER /CODE K-4/
ASST DIRECTOR FOR COMPUTATION

NATIONAL BUREAU OF STANDARDS
DATA PROCESSING SYSTEMS DIVISION
ROOM 239, BLDG. 10
ATTN A. K. SMILOW
WASH 25, D. C.

GEORGE C. FRANCIS
COMPUTING LAB, BRL
ABERDEEN PROVING GROUND, MARYLAND

OFFICE OF NAVAL RESEARCH
BRANCH OFFICE CHICAGO
230 N. MICHIGAN AVENUE
CHICAGO, ILLINOIS 60601

COMMANDING OFFICER
O N R BRANCH OFFICE
1030 E. GREEN STREET
PASADENA, CALIFORNIA

COMMANDING OFFICER
O N R BRANCH OFFICE
1000 GEARY STREET
SAN FRANCISCO 9, CALIFORNIA

WAYNE STATE UNIVERSITY
DETROIT, MICHIGAN
ATTN DEPT. OF SLAVIC LANGUAGES,
PROF. HARRY H. JOSSELSON 48202

HEBREW UNIVERSITY
JERUSALEM, ISRAEL
ATTN PROF. Y. BAR-HILLEL

NATIONAL PHYSICAL LABORATORY
TEDDINGTON, MIDDLESEX
ENGLAND
ATTN DR. A. M. UTTLEY, SUPERINTENDENT,
AUTONOMICS DIVISION

COMMANDING OFFICER
HARRY DIAMOND LABORATORIES
ATTN LIBRARY
WASHINGTON, D. C. 20438

COMMANDING OFFICER AND DIRECTOR
U. S. NAVAL TRAINING DEVICE CENTER
PORT WASHINGTON
LONG ISLAND, NEW YORK
ATTN TECHNICAL LIBRARY

DEPARTMENT OF THE ARMY
OFFICE OF THE CHIEF OF RESEARCH + DEVELOPMENT
PENTAGON, ROOM 3D442
WASHINGTON 25, D. C.
ATTN MR. L. H. GEIGER

CAMBRIDGE LANGUAGE RESEARCH UNIT
20 MILLINGTON ROAD
CAMBRIDGE, ENGLAND
ATTN MRS. MARGARET M. BRAITHWAITE

NATIONAL SECURITY AGENCY
FORT GEORGE G. MEADE, MARYLAND
ATTN LIBRARIAN, C-332

LINCOLN LABORATORY
MASSACHUSETTES INSTITUTE OF TECHNOLOGY
LEXINGTON 73, MASSACHUSETTES
ATTN LIBRARY

DR. WILLIAM S-Y WANG
DIVISION OF LINGUISTICS
OHIO STATE UNIVERSITY
COLUMBUS 14, OHIO

PROFESSOR MIEKO HAN
DEPARTMENT FOR ASIAN STUDIES
UNIVERSITY OF SOUTHERN CALIFORNIA
UNIVERSITY PARK
LOS ANGELES, CALIFORNIA 900007