ANNUAL REPORT: AUTOMATIC INDEXING AND ABSTRACTING PART II ENGLISH INDEXING OF RUSSIAN TECHNICAL TEXT

M-21-66-2 March 1966

Annual Progress Report
Office of Naval Research
Contract Nonr 4440(00)

Electronic Sciences Laboratory
Lockheed Palo Alto Research Laboratory
LOCKHEED MISSILES & SPACE COMPANY
A Group Division of Lockheed Aircraft Corporation
Palo Alto, California

PRÉCIS
RESEARCH PROGRESS REPORT

Title: "Annual Progress Report: Automatic Indexing and Abstracting, Part II. English Indexing of Russian Technical Text," H. R. Robison, Annual Progress Report, Part II, Office of Naval Research, Contract Nonr 4440(00)

Background: This investigation is concerned with the development of automatic indexing, abstracting, and extracting systems. Basic investigations in English morphology, phonetics, and syntax are pursued as necessary means to this end.

Condensed Report Contents: The following report describes a computer system for the IBM 7094 which produces English indexes of technical Russian text.

Part of the indexing system produces a machine dictionary on magnetic tape. This dictionary is a computer representation of standard English-Russian phrase technical dictionaries. A machine dictionary must exist for the same field as the text being indexed.

The indexing portion of the system operates upon the machine dictionary. Russian text phrases are matched against Russian dictionary phrases. When a match is found, the English translation is extracted from the dictionary. The final index is constructed from the set of such English translations.

The Russian dictionary entries are in canonical form. The indexing system contains reverse inflection algorithms which transform text phrases to their canonical forms.

For Further Information: The complete report is available in the major Navy technical libraries and can be obtained from the Defense Documentation Center. A few copies are available for distribution by the author.

FOREWORD

This report is Part II of the Annual Progress Report: Automatic Indexing and Abstracting submitted to the Office of Naval Research under Contract Nonr 4440(00). The work was jointly supported by the Independent Research Program of Lockheed Missiles & Space Company, and the Office of Naval Research.

ABSTRACT

The following report describes a computer system for the IBM 7094 which produces English indexes of technical Russian text.

Part of the indexing system produces a machine dictionary on magnetic tape. This dictionary is a computer representation of standard English-Russian technical phrase dictionaries.

The indexing portion of the system matches Russian text phrases against Russian dictionary phrases. Dictionary phrases are in canonical form; reverse inflection algorithms transform text phrases to their canonical form. When a match is found, the English translation of the match is extracted from the dictionary. The final index is constructed from the set of such English translations.

ACKNOWLEDGMENTS

The author would like to thank Serge Kassatkin of the Slavic Language Department, University of California, Berkeley, for his suggestions and help during the preparation of the Indexing System; Bruce Dwelley who, with the author and Mr. Kassatkin, devised the reverse inflection algorithms; and Dr. Martin Billik who formalized the indexing system in mathematical notation.

CONTENTS

Section		Page
	FOREWORD	iii
	ABSTRACT	v
	ACKNOWLEDGMENTS	vii
	ILLUSTRATIONS	xi
	TABLES	xi
1	INTRODUCTION	1-1
	1.1 General Discussion	1-1
	1.2 Translation of Phrases	1-2
	1.3 Index Entries: An Example	1-4
	1.4 The Indexing System	1-6
2	INFLECTIONAL NATURE OF THE RUSSIAN LANGUAGE	2-1
	2.1 Inflection Defined	2-1
	2.2 Inflection and Reverse Inflection	2-3
	2.3 Types of Russian Machine Dictionaries	2-4
	2.4 The Reverse Inflection Algorithm	2-5
3	PHRASE DICTIONARIES, PHRASES	3-1
	3.1 Phrase Dictionaries	3-1
	3.2 Commercially Available Phrase Dictionaries	3-1
	3.3 Examples of Some Typical Phrases	3-2
	3.4 Phrase Paradigms, Canonical Forms, Reverse Inflection	3-3
4	THE DICTIONARY CREATION PROGRAM	4-1
	4.1 General Discussion	4-1
	4.2 Selection of Representative Words and Their Coordinates	4-2
	4.3 Finding the Representative Word	4-4
	4.4 Ambigu[illegible]ngs	4-5
	4.5 A Word [illegible]djectives	4-[illegible]

Section			Page
	4.6	Representative Words, Their Coordinates and Parts-of-Speech as an Argument Function Table	4-7
	4.7	Representation of Phrases	4-10
	4.8	Logical Sums of Russian Phrases and English Translations as an Argument/Function	4-10
	4.9	Format of the Computer Dictionary	4-11
5	THE RAW INDEX		5-1
	5.1	The Algorithm	5-1
	5.2	Transformation of Text Phrase to Canonical Form	5-2
	5.3	Computer Representation of Adjectives	5-6
	5.4	Indexing Failures: How They Can Be Corrected	5-7
	5.5	Semi-Automatic Dictionary	5-8
	5.6	Example of Entry Selection for the Raw Index	5-9
6	FINAL INDEX		6-1
	6.1	General Discussion	6-1
	6.2	Simple Index	6-1
	6.3	Complex Index	6-1
7	CONCLUSIONS		7-1
	7.1	Extracting	7-1
	7.2	Translation	7-1
	7.3	Retrieval	7-1
8	REFERENCES		8-1
	8.1	Cited References	8-1
	8.2	Uncited References	8-1

Appendix

A	THE FIRST AND SECOND VERB CONFIGURATIONS	A-1
B	REVERSE INFLECTION ALGORITHM	B-1
C	COMPRESSION OF PHRASES INTO A LOGICAL SUM	C-1
D	CONCEPTUAL DESCRIPTION OF RUSSIAN TEXT PHRASE TRANSLATOR	D-1
E	OUTPUT OF THE INDEXING SYSTEM	E-1

ILLUSTRATIONS

Figure		Page
1-1	Phrase Translations	1-3
1-2	Approach to Indexing	1-5
2-1	Masculine and Neuter Genders	2-2
2-2	Feminine Gender	2-2
2-3	Declension of Adjectives	2-3

TABLES

Table		Page
4-1	Endings Used to Partition Canonical Forms Into Their Part-of-Speech Categories	4-3

Section 1
INTRODUCTION

1.1 GENERAL DISCUSSION

This report describes a computer system for the IBM 7094 which produces a deep English index of untranslated scientific Russian text. The index is printed in a back-of-the-book-type format. Though verbs may appear in a human-produced index of this type, such an index consists, for the most part, of an alphabetically arranged collection of nouns and their modifiers. The computer-produced index described here indexes nouns and their modifiers with great accuracy. If desired, the system will also perform cross indexing.

In the analysis and programming of the indexing system, the most attention was devoted to nouns and adjectives. However, the system will index verbs if desired. Only a certain noun configuration will not, at present, be indexed. This will be discussed in the body of the report.

This report has been written for readers having little, if any, knowledge of the Russian language. This fact has led to a somewhat lengthy report as Russian examples have been used widely.

Section 2 discusses the inflectional nature of Russian, and the concepts of paradigm and reverse inflection. Those familiar with Russian, even on general terms, can ignore this section.

Section 3 extends the meanings of paradigm and reverse inflection to phrases.

Sections 4 and 5 describe the construction of the machine dictionary and the Raw Index, respectively.

Section 6 describes the rules whereby the Final Index is formed from the Raw Index.

Section 7 describes possible uses to which the Indexing System may be put.

Appendix D uses a formal, compact notation to describe the formation of the Dictionary Creation Program and the Raw Index. The verbal description of the Indexer, which occupies most of the report, can be read independently of this section. On the other hand, an understanding of this section will in itself give a complete understanding of the formation of the Raw Index.

1.2 TRANSLATION OF PHRASES

It should be stressed that this system produces English indexes of untranslated Russian texts. It does so by recourse to a computer-implemented technical phrase dictionary. Such dictionaries, both Russian-English and English-Russian, are commercially available for a wide variety of technical fields. Part of the indexing system will convert any such dictionary to a machine dictionary on magnetic tape. A machine dictionary must be available for the same technical field as the corpus being indexed.

The necessity for using such dictionaries can be stated briefly: the translation of a technical phrase is a function of all the words within it taken together. Only occasionally is the translation of a phrase equal to the translation of the words composing it.

Figure 1-1 lists five Russian phrases. The word-for-word translations of the five phrases are:

- Automatic telephone system
- Cathode with open upper end
- Contact combs for recalculation
- Equally accessible bunches
- Clearance with released anchor

1. АВТОМАТИЧЕСКАЯ ТЕЛЕФОННАЯ СИСТЕМА

AUTOMATIC TELEPHONE SYSTEM

2. КАТОД С ОТКРЫТЫМ ВЕРХНИМ КОНЦОМ

OPEN CATHODE

3. КОНТАКТНЫЕ ГРЕБЕНКИ ДЛЯ ПЕРЕСЧЕТА

TRANSLATION FIELD

4. РАВНОДОСТУПНЫЕ ПУЧКИ

EQUALLY ACCESSIBLE TRUNK GROUPS

5. ЗАЗОР ПРИ ОТПУЩЕННОМ ЯКОРЕ

RELEASED GAP

Fig. 1-1 Phrase Translations

Below each Russian phrase in Fig. 1-1 appears the English translation selected from an electronics phrase dictionary. Only in the first phrase are the two translations the same.

1.3 INDEX ENTRIES: AN EXAMPLE

Before proceeding to the body of the report, a brief example of the formation of an index will be demonstrated. Figure 1-2 illustrates a sentence in which six single words and three phrases are contained in a phrase dictionary on nuclear physics.

It frequently happens that a text phrase does not have a translation in the dictionary, while the elements composing it do have dictionary translations. In such a case, all elements which are contained in the dictionary are translated and the resulting combination forms an entry for the index.*

...КАМЕРУ ВИЛЬСОНА В МАГНИТНОМ ПОЛЕ....
CLOUD CHAMBER MAGNETIC FIELD
CLOUD CHAMBER IN MAGNETIC FIELD

Sometimes elements of text phrases are high frequency words -- such as some, though not most, prepositions. The translations of such words will be incorporated into the index where possible.

Figure 1-2 shows the sentence to be indexed. The phrases occurring in the nuclear dictionary are underlined.

The index contains the following items:

- Cloud chamber in magnetic field
- Detector of particles

*It will be necessary in the report to discuss text phrases and dictionary phrases. To avoid confusion, text phrases will appear with a string of dots which represent the sentence in which the phrase appears.

В КАЧЕСТВЕ ДЕТЕКТОРА ЧАСТИЦ МЫ ПРИМЕНИЛИ КАМЕРУ ВИЛЬСОНА В МАГНИТНОМ ПОЛЕ

И ИДЕНТИФИЦИРОВАЛИ ЧАСТИЦЫ ПО ИХ ПОТЕРЯМ ЭНЕРГИИ ПОСЛЕ ПРОХОЖДЕНИЯ

ЧЕРЕЗ СВИНЦОВЫЕ ПОГЛОТИТЕЛИ.

PARTICLES CLOUD CHAMBER MAGNETIC FIELD

AS THE DETECTOR OF [FRACTIONS/PARTS] WE UTILIZED A [CELL/CHAMBER OF WILSON] IN A [MAGNETIC FIELD/MARGIN/BRIM]

PARTICLES LOSS

AND IDENTIFIED THE [FRACTIONS/PARTS] BY THEIR [LOSS/WASTE] OF ENERGY AFTER PASSAGE THROUGH LEAD ABSORBERS.

Fig. 1-2 Approach to Indexing

- Energy loss after passage through lead absorbers
- Lead absorbers, energy loss after passage through
- Magnetic field, cloud chamber in

1.4 THE INDEXING SYSTEM

The indexing system consists of two main parts, the Dictionary Creation Program (DCP), and the Indexer.

1.4.1 Dictionary Creation Program

The DCP creates a machine dictionary which is used by the Indexer. Creation of the machine dictionary takes place prior to and independently of the Indexer's operation. The dictionary itself is stored on magnetic tape. Once a dictionary for a given field or subfield of science has been created, it is used by the Indexer to index all texts in the same field.

1.4.2 The Indexer

The operation of the Indexer proceeds in two steps: creation of a Raw Index, and creation, from the Raw Index, of the Final Index.

- Raw Index
  Phrases in the Russian text which is being indexed are matched against Russian phrases in the machine dictionary. When a match is found, the English translation of the Russian phrase is retrieved from the dictionary and placed in the Raw Index, along with information regarding the position and length of the phrase on the page, and the part-of-speech of the main word in the phrase.
- Final Index
  The Raw Index is examined and, using the information contained in it, a Final Index is constructed.

Section 2
INFLECTIONAL NATURE OF THE RUSSIAN LANGUAGE

2.1 INFLECTION DEFINED

Before discussing the indexing system in detail, it is essential to understand something of the inflection problem in Russian.

Inflection is that property of a language by which a particular relationship between two words, usually nouns or pronouns, is expressed by a change in form of one of the words. This type of relationship is usually referred to as case.

English was once a more highly inflected language than at present, but there still remains a residue of the old case structure. Thus, the pronoun I inflexts to me when it becomes the direct object of a verb or the object of a preposition, as who changes to whom in the same circumstances.

In Russian there are six cases, and two numbers – singular and plural. In Russian, therefore, a noun may have as many as twelve forms. Figures 2-1 and 2-2 show the case forms for the three genders of Russian nouns – masculine, feminine, and neuter. The names of the six cases are nominative, genitive, dative, accusative, instrumental, and prepositional.

Russian adjectives also inflect according to the case of the noun they modify (Fig. 2-3).

Figures 2-1 through 2-3 illustrate the declensions of regular nouns and adjectives. The set of all inflected forms of a given word is called the paradigm of the word. Generally, some of the members of the paradigm coincide. The set which contains only distinct entries is called the reduced paradigm.

	Singular					
	Masculine			*Neuter*		
	Hard	*Soft*	*Soft*	*Hard*	*Soft*	*Soft*
Nom.	стол	музе́й	дождь	ме́сто	по́ле	зда́ние
Gen.	стола́	музе́я	дождя́	ме́ста	по́ля	зда́ния
Dat.	столу́	музе́ю	дождю́	ме́сту	по́лю	зда́нию
Acc.	стол	музе́й	дождь	ме́сто	по́ле	зда́ние
Instr.	столо́м	музе́ем	дождём	ме́стом	по́лем	зда́нием
Prep.	столе́	музе́е	дожде́	ме́сте	по́ле	зда́нии
	Plural					
Nom.	столы́	музе́и	дожди́	места́	поля́	зда́ния
Gen.	столо́в	музе́ев	дождей	мест	поле́й	зда́ний
Dat.	стола́м	музе́ям	дождя́м	места́м	поля́м	зда́ниям
Acc.	столы́	музе́и	дожди́	места́	поля́	зда́ния
Instr.	стола́ми	музе́ями	дождя́ми	места́ми	поля́ми	зда́ниями
Prep.	стола́х	музе́ях	дождя́х	места́х	поля́х	зда́ниях

Fig. 2-1 Masculine and Neuter Genders

	Singular			
	Hard	*Soft*	*Soft*	*Soft*
Nom.	ко́мната	неде́ля	дверь	фами́лия
Gen.	ко́мнаты	неде́ли	две́ри	фами́лии
Dat.	ко́мнате	неде́ле	две́ри	фами́лии
Acc.	ко́мнату	неде́лю	дверь	фами́лию
Instr.	ко́мнатой (ою)	неде́лей(ею)	две́рью	фами́лией(ею)
Prep.	ко́мнате	неде́ле	две́ри	фами́лии
	Plural			
Nom.	ко́мнаты	неде́ли	две́ри	фами́лии
Gen.	ко́мнат	неде́ль	дверей	фами́лий
Dat.	ко́мнатам	неде́лям	дверям	фами́лиям
Acc.	ко́мнаты	неде́ли	две́ри	фами́лии
Instr.	ко́мнатами	неде́лями	дверями	фами́лиями
Prep.	ко́мнатах	неде́лях	дверях	фами́лиях

Fig. 2-2 Feminine Gender

Hard: -'ый (-'ий); -о́й

	Masculine	*Neuter*	*Feminine*	*Plural All Genders*
Nom.	но́вый	но́вое	но́вая	но́вые
Gen.	но́вого	но́вого	но́вой	но́вых
Dat.	но́вому	но́вому	но́вой	но́вым
Acc.	но́вый(ого)	но́вое	но́вую	но́вые(ых)
Instr.	но́вым	но́вым	но́вой (ою)	но́выми
Prep.	но́вом	но́вом	но́вой	но́вых

Soft: -'ий

	Masculine	*Neuter*	*Feminine*	*Plural All Genders*
Nom.	си́ний	си́нее	си́няя	си́ние
Gen.	си́него	си́него	си́ней	си́них
Dat.	си́нему	си́нему	си́ней	си́ним
Acc.	си́ний(его)	си́нее	си́нюю	си́ние(их)
Instr.	си́ним	си́ним	си́ней (ею)	си́ними
Prep.	си́нем	си́нем	си́ней	си́них

Fig. 2-3 Declension of Adjectives

There are, in addition, a great number of irregular nouns and adjectives whose inflections differ in varying degree from those shown. We are not concerned at present, however, with details of the inflectional structure but rather with its size and nature. The important fact shown in Figs. 2-1 through 2-3 is that, despite the coincidence of two or more elements of a paradigm, there remains an overpowering multiplicity of possible forms.

(This report is concerned mainly with adjectives and nouns. Nevertheless, since verbs will be indexed if desired they will be discussed where it seems pertinent to do so. Appendix A contains a listing of verbal forms for the first and second conjugations.)

2.2 INFLECTION AND REVERSE INFLECTION

In manual dictionaries the paradigm of a word is represented by a single element of the paradigm. This element is called the canonical form of the paradigm. In an English

dictionary cathode is the canonical form of the paradigm {cathode, cathodes}. Dictionary makers assume the ability on the user's part to transform an inflected form of the word to its canonical form.

It goes without saying that there exist in both English and Russian many "irregular" words, so called because the "regular" transformations will not suffice to transform the canonical form into elements of its paradigm and vice-versa. Nevertheless, the very existence of canonical dictionaries implies that the great majority of words in them are susceptible to "regular" transformations.

The process of transforming a canonical form into one or more elements of its paradigm will be called inflection. The inverse transformation, deriving the canonical form from an element of the reduced paradigm, will be called reverse inflection.

2.3 TYPES OF RUSSIAN MACHINE DICTIONARIES

There are two extreme forms which an automatic dictionary may take:

- An inflected or paradigm dictionary in which each element of each paradigm is represented by a distinct entry
- A canonical dictionary in which a paradigm is represented by a single entry – its canonical form

The paradigm dictionary operates with a simple table-look-up program, but it has, of course, the disadvantage of great size.

A canonical dictionary implies the existence of an algorithm – of some complexity – which can perform the reverse inflection transformation on words encountered in text. Hybrids of these two extremes are, of course, possible.

2.4 THE REVERSE INFLECTION ALGORITHM

Several years ago a reverse inflection algorithm was developed and programmed for use in a Russian parsing program being developed at Lockheed. The algorithm itself is the subject of a forthcoming report. but it is discussed in general terms in Appendix B.

The dictionary which is used in conjunction with the reverse inflection algorithm is a dictionary of canonical forms; the canonical forms are the classically accepted ones – nominative singular for nouns, nominative singular masculine gender for adjectives. and the infinitive for verbs and participles.

The principle of the algorithm can be stated briefly: potential canonical forms of a given text word are constructed by removing certain terminal strings of letters and then adding new terminal strings. After each potential canonical form is constructed. an attempt is made to find it in the dictionary. If the potential canonical form has no match in the dictionary. a new potential form is constructed and so on. until a true canonical form is constructed or until all possible constructions for the word in question have been exhausted.

At the time. then. that the indexing program was written. there existed a reverse inflection algorithm designed to operate on a canonical dictionary whose entries are classically defined canonical forms. The dictionary entries are single words. and the reverse inflection algorithm operates on a single text word at a time. Let us define such a dictionary as D_s and the reverse inflection algorithm as R_s. We will now extend the definitions of this section to a phrase dictionary.

Section 3

PHRASE DICTIONARIES, PHRASES

3.1 PHRASE DICTIONARIES

In the Introduction it was stated that the Indexer uses a dictionary of technical phrases to compute the index. These dictionaries cover the terminology of a given field or subfield of science and thus provide a list of phrase descriptors for the field in question.

This is a different type of dictionary than the ordinary dictionary whose entries are single words. We are dealing now with a dictionary whose entries are phrases.

3.2 COMMERCIALLY AVAILABLE PHRASE DICTIONARIES

English-Russian technical phrase dictionaries are lists of English phrases and their Russian translations. Single words may occur, but, in general, these dictionaries are phrase dictionaries.

A few dictionaries of this type have been published in the United States (where, of course, the format is usually Russian-English), but for the most part they have been compiled in the Soviet Union. (References 1 through 6.)

Most Soviet dictionaries, having been prepared for translation of English to Russian, are arranged alphabetically according to the English alphabet.

3.3 EXAMPLES OF SOME TYPICAL PHRASES

The following is a list of phrases taken from a nuclear physics dictionary

ПОЛНАЯ МОЩНОСТЬ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА

TOTAL REACTION POWER DENSITY

ОРБИТАЛЬНАЯ ПЛОСКОСТЬ

ORBITAL PLANE

ФОТОРОЖДЕНИЕ

PHOTOPRODUCTION

ЭФФЕКТИВНОЕ СЕЧЕНИЕ ДЛЯ ДЕЛЕНИЯ УРАНА

CROSS SECTION FOR URANIUM FISSION

ЭФФЕКТ ПЕРЕНОСА

TRANSFER EFFECT

КОСМИЧЕСКИЙ

COSMIC

ИНДУКТИРОВАТЬ

INDUCE

ДИФФЕРЕНЦИРУЮЩАЯ СХЕМА

DIFFERENTIATING NETWORK

НАНОСИТЬ В ЗАВИСИМОСТИ ОТ

PLOT AGAINST

3.4 PHRASE PARADIGMS, CANONICAL FORMS, REVERSE INFLECTION

The definitions of Section 2 can easily be extended to include phrases. Consider, for example, the first phrase in the above set of examples. Its paradigm is:

Singular	Nom.	ПОЛНАЯ МОЩНОСТЬ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
	Gen.	ПОЛНОЙ МОЩНОСТИ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
	Dat.	ПОЛНОЙ МОЩНОСТИ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
	Acc.	ПОЛНУЮ МОЩНОСТЬ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
	Instr.	ПОЛНОЙ МОЩНОСТЬЮ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
	Prep.	ПОЛНОЙ МОЩНОСТИ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
Plural	Nom.	ПОЛНЫЕ МОЩНОСТИ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
	Gen.	ПОЛНЫХ МОЩНОСТЕЙ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
	Dat.	ПОЛНЫМ МОЩНОСТЯМ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
	Acc.	ПОЛНЫЕ МОЩНОСТИ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
	Instr.	ПОЛНЫМИ МОЩНОСТЯМИ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
	Prep.	ПОЛНЫХ МОЩНОСТЯХ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА

The reduced paradigm is:

Singular	Nom.	ПОЛНАЯ МОЩНОСТЬ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
	Gen., Dat., Prep.	ПОЛНОЙ МОЩНОСТИ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
	Acc.	ПОЛНУЮ МОЩНОСТЬ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
	Instr.	ПОЛНОЙ МОЩНОСТЬЮ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
Plural	Nom., Acc.	ПОЛНЫЕ МОЩНОСТИ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
	Gen.	ПОЛНЫХ МОЩНОСТЕЙ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
	Dat.	ПОЛНЫМ МОЩНОСТЯМ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
	Instr.	ПОЛНЫМИ МОЩНОСТЯМИ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА
	Prep.	ПОЛНЫХ МОЩНОСТЯХ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА

The cases listed on the left refer to the case of the leftmost noun. We will define the canonical form of the phrase as the form whose leftmost noun is in the nominative singular case. Thus in this example

ПОЛНАЯ МОЩНОСТЬ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА

is the canonical form of the phrase paradigm. The process of transforming the canonical form of a phrase into one or more elements of its paradigm is called inflection and, as before, the inverse transformation deriving the canonical form of a phrase from an element of the paradigm is called reverse inflection. We will denote by D_p a dictionary whose entries are canonical phrases. R_p is the reverse inflection algorithm that transforms members of the phrase paradigm to canonical form.

Section 4
THE DICTIONARY CREATION PROGRAM

4.1 GENERAL DISCUSSION

An examination of the Russian phrases of a phrase dictionary indicates that each phrase has within it a main or pivotal word upon which the rest of the phrase, so to speak, depends. Usually this word is the leftmost noun or verb of the phrase. (If the phrase consists of a single word, then obviously this is the pivotal word.) This pivotal word occurs in the canonical form—infinitive of verbs, nominative singular for nouns.

We will see that this word may automatically be identified with great accuracy. This main or pivotal word will be referred to from now on as the representative word.

The position of the representative word within the dictionary phrase is defined as the left and right limits of the phrase with respect to the representative word. These left and right limits are called coordinates.

The DCP selects a set of representative words from the phrases of the phrase dictionary. If a word of this set or an element of the paradigms of this set occurs in text, it is a signal to the Indexer that this text word may be the representative of an entire text phrase, the equivalent of which is contained in the machine dictionary.

Next, the Indexer retrieves the coordinates of the representative word and uses these coordinates to examine the text environment of that word which was originally transformed into the representative word.

In short, the construction of the set of representative words is equivalent to the establishment of a set of signals to inform the Indexer that the environment of a given text

word must be examined in detail because it may contain a phrase whose equivalent occurs in the machine dictionary. In addition to the set of representative words, the DCP creates a set of Russian phrases and a set of English translations of these phrases. Links between these three sets are also, of course, established so that the Indexer may thread its way from representative word to Russian phrase represented by that word to English translation of the phrase.

4.2 SELECTION OF REPRESENTATIVE WORDS AND THEIR COORDINATES

The dictionary creation program examines the string of Russian words making up the Russian portion of a dictionary entry and

- Selects the representative word of the phrase
- Assigns a part-of-speech category of noun, verb, or adjective* to the representative word
- Determines the coordinates of the representative word within the phrase

The representative word is defined as follows:

- The leftmost noun or verb in the phrase
- If there is no noun or verb in the phrase, the phrase is an adjective (or a string of adjectives); the rightmost adjective is selected as the representative word

These rules imply an ability to distinguish between adjectives, nouns, and verbs. Since the Dictionary Creation program does not have recourse to a dictionary, is in

*Participles in phrase dictionaries generally behave syntactically as though they were attributive adjectives (Appendix A). The canonical form of a participle is generally considered the infinitive from which the participle derives. We will see that adjectives which lie to the left of representative words or are representative words themselves are listed in the machine dictionary by their stems. This is true of participles as well, the form being the word minus its adjectival ending. If -СЯ or -Сь had to be removed from the participle to get to its adjectival ending then -СЯ or -Сь is restored to the participle stem. In summary, participles behave like attributive adjectives and are handled as such by the indexing system. Each time the word "adjective" occurs in the report it can be read as "adjective and/or participle."

fact actually making the dictionary, it is evident that the separation must be made on the basis of the actual words occurring in the phrases. Table 4-1 shows the endings used to partition canonical forms into their part-of-speech categories.

Table 4-1

ENDINGS USED TO PARTITION CANONICAL FORMS INTO THEIR PART-OF-SPEECH CATEGORIES

Verb-Part[(a)]	Adj	Noun	Adj-Noun	Verb
•СЯ	•ЖИЙ	•ОСТЬ	•НИЙ	•ТЬ
•СЬ	•ЧИЙ	•ОЧЬ	•ОЙ	•ТИ
	•ШИЙ	•ЕЧЬ	•ЕЕ	
	•ЩИЙ			
	•ГИЙ			
	•КИЙ			
	•ХИЙ			
	•ЫЙ			
	•ЯЯ			
	•АЯ			
	•ОЕ			
	•ЖИЕ			
	•ЧИЕ			
	•ШИЕ			
	•ЩИЕ			
	•ГИЕ			
	•КИЕ			
	•ХИЕ			
	•ЫЕ			

(a) If these endings occur, remove them before searching for other endings. Restore them after suffix examination.

The word endings are examined in conjunction with each other. Thus the word ПЕРИОДИЧНОСТЬ is not a verb but a noun because the verbal (•ТЬ) ending is contained in a larger noun (•ОСТЬ) ending. The larger ending takes precedence.

If a word has none of the above endings it is called a noun. Thus СЕЧЕНИЕ is a noun because it has no ending in the above categories. The endings ОЧЬ and •ЕЧЬ

are verbal endings as well as noun endings, but the verbs form such a small class of technical words that it was decided to designate words with such endings as nouns. НАНОСИТЬ is called a verb only after determining that its verbal (-ТЬ) ending is not contained in the longer (-ОСТЬ) noun ending. ПРЯМОЙ is called both a noun and an adjective because its ending is ambiguous.

4.3 FINDING THE REPRESENTATIVE WORD

The separation is made on the basis of endings. The method is quite accurate though not 100 percent accurate. It must also be kept in mind that the method described below applies to Russian dictionary phrases in canonical form, not to other members of the canonical form's paradigm.

Definition: A terminator is an unambiguous noun or verb.

Rule 1: Scan Russian phrase from left to right. The scan halts at the first terminator. This terminator is the representative word.

Rule 2: The Russian phrase begins with an ambiguous word or a string of ambiguous words. Three possibilities arise and are handled as follows:

(a) Call the first word a noun and apply Rule 1.

(b) The ambiguous string has k words. Call each of the k words adjectives. If there is a k + 1 word call it a noun and apply Rule 1. If there is no k + 1 word apply Rule 3.

Rule 3: No terminator is encountered and no ambiguous words are encountered. The phrase consists of an adjective or a string of adjectives. The scan is terminated by the last word. Call the last word the representative word.

The practical result of Rule 2 is that two representative words are selected, two phrases, and two linkages to the English translation. Thus, a word ending in -ОЙ is listed two times as a representative word, once as a noun, once as an adjective. (Examples are shown in subsection 4.4.)

In summary then, the representative word is defined by means of the three rules as:

- The first noun or verb encountered
- If no noun and no verb encountered, then the representative word is the last adjective of the leading adjective string

The phrases previously listed are repeated below with the representative words underlined. The coordinates of the representative word are easily computed once the representative word itself has been determined. The coordinates are listed in the right hand column. Coordinates of (0.0) indicate that the phrase consists of a single word, namely the representative word itself.

ПОЛНАЯ МОЩНОСТЬ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА	(1.4)
ОРБИТАЛЬНАЯ ПЛОСКОСТЬ	(1.0)
ФОТОРОЖДЕНИЕ	(0.0)
ЭФФЕКТИВНОЕ СЕЧЕНИЕ ДЛЯ ДЕЛЕНИЯ УРАНА	(1.3)
ЭФФЕКТ ПЕРЕНОСА	(0.1)
КОСМИЧЕСКИЙ	(0.0)
ИНДУКТИРОВАТЬ	(0.0)
ДИФФЕРЕНЦИРУЮЩАЯ СХЕМА	(1.0)
НАНОСИТЬ В ЗАВИСИМОСТИ ОТ	(0.3)

4.4 AMBIGUOUS ENDINGS

Phrases containing **•ОЙ , НИЙ** and **ЕЕ** are handled differently, as has been stated before.

ПРЯМОЙ ПОТОК STRAIGHT-THROUGH FLOW	**ПРЯМОЙ ПОТОК** NOUN	1
	ПРЯМОЙ ПОТОК ADJ NOUN	2
ВИХРЕВОЙ ТОК EDDY CURRENT	**ВИХРЕВОЙ ТОК** NOUN	3
	ВИХРЕВОЙ ТОК ADJ NOUN	4
ПОЛОНИЙ POLONIUM	**ПОЛОНИЙ** NOUN	5
	ПОЛОНИЙ ADJ	6

But

КАНАЛ АКТИВНОЙ ЗОНЫ CORE CHANNEL	**КАНАЛ АКТИВНОЙ ЗОНЫ**

because the program recognizes **КАНАЛ** as a noun, hence the representative word before it encounters **АКТИВНОЙ** . The assignment of the correct part-of-speech to the representative word is important; the indexing program uses this syntactic information in constructing the final index. The representative words in items 1, 3, and 6, it should be noted, have been assigned the incorrect part-of-speech. It will be shown in subsection 5.3 how this error is handled by the system. Let us simply note for now that the error is unavoidable because the program cannot determine if an **ОЙ , НИЙ** or **ЕЕ** word is a noun or an adjective; therefore it regards these endings as ambiguous

1.5 A WORD ABOUT ADJECTIVES

Unlike nouns and verbs, an adjective, if it is a representative word, is carried in the dictionary not in its classical canonical form, but by its stem. In addition, if a representative word of a phrase is a noun, then adjectives to the left of the representative word have had their endings removed. The reason for this is discussed in subsection 5.2.3 where the discussion arises naturally as part of the indexing procedure.

4.6 REPRESENTATIVE WORDS, THEIR COORDINATES AND PARTS-OF-SPEECH AS AN ARGUMENT/FUNCTION TABLE

After the DCP has determined for each entry the representative word and the part-of-speech and coordinates of the representative word, it then proceeds to construct an argument/function table where representative words make up the argument portion of the table and coordinates and parts-of-speech the function portion.

The argument table, of course, can contain only distinct entries. Now if a particular representative word occurs more than once, Rules 1, 2, and 3 will assure that the word in each occurrence has the same part-of-speech. The coordinates of each word, however, may differ. Therefore, all sets of coordinates belonging to identical representative words must be examined so that only the largest coordinates will be entered in the function portion of the table.

Now when the Indexer finds a representative word in text and retrieves its function, it knows not only the representative word's part-of-speech, but also the size (i.e., the number of words, expressed by the coordinate values) of the longest phrases which it represents.

In text, of course, phrases smaller than the maximum may occur. Thus, the Indexer must construct and seek, for any given representative word, all possible text phrases that surround it up to the maximum size permitted by the function coordinates.

Example 1: The following phrases, having the same representative word, occur in a dictionary.

(1) method/**МЕТОД** (0, 0)
(2) curve fitting method/**МЕТОД ПОДГОНКИ КРИВОЙ** (0, 2)
(3) activation method/**АКТИВАЦИОННЫЙ МЕТОД** (1, 0)
(4) radiation prospecting method/**РАДИАЦИОННЫЙ МЕТОД ПОИСКОВ** (1, 1)

Argument	Function
МЕТОД	noun, (1, 2)

The coordinates (1, 2) are the left and right limits of the largest possible phrases whose representative word is **МЕТОД** .

If, in the following text sequence, w_3 is the representative word with coordinates, say, (1, 2)

$$w_1 w_2 \underline{w_3} w_4 w_5 w_6 \cdots$$

then the possible sets of text phrases based on w_3 and coordinates (1, 2) are

$\cdots w_2 w_3 w_4 w_5 \cdots$	(1, 2)
$\cdots w_2 w_3 w_4 \cdots$	(1, 1)
$\cdots w_3 w_4 w_5 \cdots$	(0, 2)
$\cdots w_2 w_3 \cdots$	(1, 0)
$\cdots w_3 w_4 \cdots$	(0, 1)
$\cdots w_3 \cdots$	(0, 0)

In the first two phrases the representative word, w_3 , is embedded in the phrase while in the remaining four phrases w_3 is the left or right limit of the phrase.

Operationally it is useful to distinguish these two types of coordinates – those representing an embedded representative word and those representing a representative word at the phrase's limits. The reason for this distinction will be made clear by the next example. Suppose the following two dictionary phrases have the same representative word.

Example 2:

(1) powder-diffraction method/ **МЕТОД ДИФРАКЦИИ НА ПОРОШКЕ** (0, 3)

(2) radiation prospecting method/ **РАДИАЦИОННЫЙ МЕТОД ПОИСКОВ** (1, 1)

If we do not distinguish between the two kinds of coordinates we get, as before

Argument	Function
МЕТОД	noun, (1,3)

and again if w_3 = **МЕТОД** in the sequence $w_1w_2w_3w_4$ .... the possible text phrases are

... $w_2w_3w_4w_5w_6$	(1.3)
... $w_2w_3w_4w_5$	(1.2)
... $w_2w_3w_4$	(1.1)
... w_2w_3	(1.0)
... $w_3w_4w_5w_6$	(0.3)
... $w_3w_4w_5$	(0.2)
... w_3w_4	(0.1)
... w_3	(0.0)

The first two phrases [coordinates (1.3) and (1.2)] are combinations which do not occur in the dictionary and which need never be formed by the Indexer if the functions (0.3) and (1.1) are kept separate.

Therefore, the argument/function pairs are

Argument	Function
МЕТОД	noun, (0.3), (1.1)

leading to the following text phrases – three less than in the previous case where coordinates were simply merged.

$\ldots\ w_2 w_3 w_4\ \ldots\ .$ (1, 1)

$\ldots\ w_3 w_4 w_5 w_6\ \ldots\ .$ (0, 3)

$\ldots\ w_3 w_4 w_5\ \ldots\ .$ (0, 2)

$\ldots\ w_3 w_4\ \ldots\ .$ (0, 1)

$\ldots\ w_3\ \ldots\ .$ (0, 0)

The Indexer forms, and searches the dictionary for, the largest phrases first. When a text phrase is found to have a match in the dictionary, the Indexer terminates potential phrase construction based on the particular representative word it is operating with. Since the largest phrases are found first, this procedure ensures against the double indexing of nested phrases. Thus, "curve fitting method" is the indexed item, not "method."

4.7 REPRESENTATION OF PHRASES

We have discussed selection of the representative word. We have also discussed coordinates and have shown how coordinates of different phrases having the same representative word are merged. Now we will show how the Russian phrase itself is represented in the machine dictionary.

It would be extremely cumbersome to work with an actual list of Russian phrases, trying to match elements of such a list with another list computed from the text. The Russian phrase as a phrase – as a string, that is, of words – does not occur in the machine dictionary. It is represented instead by its logical sum. This means of argument compression represents the phrase as a binary number occupying one machine word. Appendix C shows how logical sums are formed.

4.8 LOGICAL SUMS OF RUSSIAN PHRASES AND ENGLISH TRANSLATIONS AS AN ARGUMENT/FUNCTION TABLE

This argument/function table consists of the logical sums of Russian phrases in the argument portion and the locations of the English translations in the function portion.

The sums are arranged in ascending order so that they are susceptible to a binary search.

It is possible, though certainly not probable, that two distinct phrases could produce identical logical sums. A test case of 8,500 English words and phrases produced only 50 duplicates and most of these were caused by two-word phrases in inverse order. For example, index arithmetic and arithmetic index produced the same logical sum.

It is not known whether such a compression device has ever been used with Russian phrases.

4.9 FORMAT OF THE COMPUTER DICTIONARY

We are now in a position to describe the structure of the machine dictionary.

First, for each letter of the English alphabet occurring in the manual dictionary, three files of information are created as follows:

FILE A

An argument/function table. The arguments are representative words. The functions are the parts-of-speech and the coordinates of the representative words.

FILE B

Also an argument/function table. The arguments are the logical sums of Russian phrases. The functions are the location in FILE C of the English translations of those Russian phrases whose logical sums are contained in the argument portion of this file.

FILE C

Not an argument/function table. Simply a list of English phrases, the Russian equivalents of which have been logically summed and stored in the argument portion of FILE B.

This threefold structure exists for each letter of the English alphabet which occurred in the phrase dictionary. If 25 letters appeared there will be 75 files.

A schematic picture of the dictionary is as follows where FILE A (A) means the A file for the English letter A.

FILE A (A)
FILE B (A)
FILE A (B)
FILE B (B)
FILE A (C)
FILE B (C)
⋮
FILE A (Z)
FILE B (Z)
FILE C (A)
FILE C (B)
FILE C (C)
⋮
FILE C (Z)

Except for a coming discussion in subsection 5.2.3, where the reasons for carrying certain adjectives in a stem form are discussed, the dictionary system used by the Indexer has now been completely described.

In summary, the dictionary system is actually composed of two distinct types of dictionary (Files A and B) and an English buffer (File C). R_s, the reverse inflection algorithm for words, and R_p, the reverse inflection algorithm for phrases, operate in conjunction with Files A and B, respectively, because the argument portions of Files A and B correspond to D_s (canonical dictionary of words) and D_p (canonical dictionary of phrases), respectively.

Section 5
THE RAW INDEX

5.1 THE ALGORITHM

The algorithm which produces English translations for intervals of ordered Russian words (i.e., the Raw Index) is described in Appendix D by means of a mathematical notation which can be read as a flow chart. The algortihm will now be described verbally in conjunction with an example.

- A page of text is read into the computer's core.
- A text word w_i is successfully transformed by R_s into its canonical form (i is the position of the word on the page). This can only happen if the canonical form of w_i , $\bar{w}$, is a representative word (i.e., if it is in the set of A files).
- The Indexer retrieves $\bar{w}$'s coordinates (ℓ, r) and part-of-speech. With the coordinates the Indexer computes the set of phrases $w_{i-k} \cdots \bar{w}\, w_{l+1} \cdots w_{i+m}$ $(\ell \leq k \leq o \ , \ o \leq m \leq r)$. Note that w's canonical form, <u>not</u> w itself occurs in each element of the phrase set.
- For each phrase in the set of phrases, a logical sum is computed (Appendix C). Then for each sum (starting with the one representing the longest word-string), the Indexer seeks a match in that B file corresponding to the A file where $\bar{w}$ was located.
- If a match is found, the Indexer has in fact transformed a text phrase to its canonical form. (See subsection 5.2 for details of this transformation.) Now the logical sum is simply a Russian phrase in a compressed form. The English equivalent of this Russian phrase is now retrieved from FILE C corresponding to FILE B where the sum match was found.

- The English translation, the part-of-speech of the Russian representative word, and the left and right sentence limits of the successfully transformed Russian phrase are now stored in a Raw Index Matrix in a row corresponding to the position of w on the original page. In addition, Russian prepositions, conjunctions, and other high-frequency words are stored in the Raw Index Matrix in rows corresponding to their positions on the page.
- The Final Index (Section 6) is constructed from elements of the Raw Index Matrix.

5.2 TRANSFORMATION OF TEXT PHRASE TO CANONICAL FORM

The selection of a representative word from a phrase and the computing of its coordinates can be considered a device for determining phrase limits in text, for determining, that is, text phrases to be used as input to R_p.

Now we will discuss the transformation of the text phrase to its canonical form. We will discuss, that is, the reverse inflection algorithm for phrases, R_p.

First, let us assume that the text phrase which has been isolated is a member of some paradigm whose canonical form is contained in the dictionary. (If this is not the case, then R_p will fail and the Indexer will pass on to the next potential text phrase.)

The text phrase and the dictionary phrase contain, at most, three components:

- The representative word
- The word string following the representative word
- The word string preceding the representative word

Either word string may be empty, but, in the general form considered here, both exist. The transformation of the text phrase to its canonical form can be considered as the transformation of each text component to its dictionary component.

5.2.1 Transformation of Representative Word

The text representative word has already been transformed to its canonical form by R_s. This was necessary to obtain the phrase coordinates in the first place.

5.2.2 Words Following Representative Word

These words do not have to be transformed. An examination of the list on page 3-2 shows that the configuration of these words with respect to each other and with respect to the representative word is fixed. Further, these words are contiguous (as are all the words in the phrase, for that matter). Finally, the cases of these words are fixed. Thus, the words following the representative word occur in text exactly as they occur in the canonical form of the phrase.

5.2.3 Words Preceding Representative Word

The only portion of a phrase which may inflect according to its use within the sentence is the representative word and words preceding it.

The representative word has been transformed by R_s. Now we will examine the preceding words.

We stated in Section 2 that the inverse inflection algorithm, R_s, transformed verbs to the infinitive, nouns to the nominative singular, and adjectives to the masculine nominative singular.

We will now show that transforming text adjectives to the masculine nominative singular will not necessarily lead to the proper dictionary phrase, and that, consequently, the adjective-transforming routine of R_s must be slightly altered.

Suppose the following phrase is in the computer dictionary.

(1) ЭФФЕКТИВНОЕ СЕЧЕНИЕ / EFFECTIVE CROSS-SECTION

СЕЧЕНИЕ , the representative word, is underlined. Now suppose the following phrase occurs in text in the instrumental case

(2) **...ЭФФЕКТИВНЫМ СЕЧЕНИЕМ....**

which the reverse inflection algorithm, R_s , operating on each word transforms* to

(3) **...ЭФФЕКТИВНЫЙ СЕЧЕНИЕ....**

Phrase 3 is not the same as phrase 1, the dictionary phrase. (In fact, phrase 3 is grammatically incorrect – a masculine adjective modifying a neuter noun.) In any event, the English translation "effective cross-section" would not be regarded as a potential indexable phrase because the Russian equivalent is being incorrectly constructed from text.

The difficulty here arises because adjectives agree in case, number, and gender with the nouns they modify. In a dictionary entry an adjective modifying a neuter or feminine noun is itself in the neuter or feminine nominative form. But the reverse inflection algorithm, R_s , operating on text words transforms all adjectives to the masculine singular nominative form. The result is that the Indexer constructs an incorrect hybrid noun unit – the adjective in the masculine singular nominative and the noun in the feminine or neuter singular nominative. A correct match with the noun phrase contained in the dictionary cannot, of course, be made. This difficulty has been resolved by altering the procedure described in subsection 4.3. Under this alteration, adjectives preceding a representative word which is a noun have their adjectival endings removed. If the phrase consists of an adjective or a string of adjectives, then the adjectival endings are removed from each word. This means that by the time the Dictionary Creation Program has found the representative word,

---

*We must assume for this example that **ЭФФЕКТИВНЫЙ** is in the dictionary. If it were not, then R_s could not transform the instrumental text form to the nominative (dictionary) form.

any adjectives preceding it have had their endings removed. If the representative word is itself an adjective, then its ending also is removed. Thus the phrase **ЭФФЕКТИВНОЕ СЕЧЕНИЕ** is contained in the dictionary as

ЭФФЕКТИВН СЕЧЕНИЕ

The Indexer, when it creates phrases from text, removes the same endings from the proper adjectives.

The reverse inflection algorithm (Appendix B) is so structured that when it correctly creates a canonical form it knows the part-of-speech of that form. Now, if the Indexer transforms a text word to a canonical form, and that form is a noun, and its coordinates indicate a left limit other than zero, then the Indexer, when scanning left in text to form the text phrase, must remove adjectival endings from words lying to the left of the text representative word. Thus

...ЭФФЕКТИВНЫМ СЕЧЕНИЕМ....

becomes

...ЭФФЕКТИВН СЕЧЕНИЕ....

which is precisely the way the phrase occurs in the dictionary. Note that **ЭФФЕКТИВН** need not appear in the dictionary as a representative word, because it is contained in the dictionary in the logical sum representing the phrase **ЭФФЕКТИВН СЕЧЕНИЕ**. The assumption here, of course, is that an adjective string lying to the left of a noun and contiguous with it modifies the noun - a dangerous assumption in literary Russian, but safe enough for scientific text.

5.3 COMPUTER REPRESENTATION OF ADJECTIVES

The phrases of page 3-2 are repeated below. Adjectival endings have been removed from adjectives preceding the representative word.

ПОЛН МОЩНОСТЬ РЕАКЦИИ НА ЕДИНИЦУ ОБЪЕМА

ОРБИТАЛЬН ПЛОСКОСТЬ

ФОТОРОЖДЕНИЕ

ЭФФЕКТИВН СЕЧЕНИЕ ДЛЯ ДЕЛЕНИЯ УРАНА

ЭФФЕКТ ПЕРЕНОСА

КОСМИЧЕС

ИНДУКТИРОВАТЬ

ДИФФЕРЕНЦИРУЮЩ СХЕМА

НАНОСИТЬ В ЗАВИСИМОСТИ ОТ

Double dictionary entries are constructed for phrases beginning with a string of ambiguous words:

ПРЯМОЙ ПОТОК
STRAIGHT-THROUGH FLOW

- ПРЯМОЙ ПОТОК — NOUN
- ПРЯМ ПОТОК — ADJ NOUN

ВИХРЕВОЙ ТОК
EDDY CURRENT

- ВИХРЕВОЙ ТОК — NOUN
- ВИХРЕВ ТОК — ADJ NOUN

ПОЛОНИЙ
POLONIUM

- ПОЛОНИЙ — NOUN
- ПОЛО — ADJ

We can now see more clearly the reason for creating double entries for **ОЙ , ЕЕ ,** and **НИЙ** phrases. One of the two representative words – the form with the adjectival ending – will never be located in the dictionary because if it or a member of its paradigm occurs in text, the adjectival ending will be removed prior to the dictionary search. And removing the adjectival ending leads to the correct (i.e., the dictionary) representation of the adjective. The incorrect representative word is thus a "wasted" entry, the price paid for making the DCP automatic.

5.4 INDEXING FAILURES: HOW THEY CAN BE CORRECTED

Certain configurations of text words will not be indexed properly. It is not believed that these types of configuration occur often enough to be considered a serious problem but they must be mentioned for completeness. In any case, they can be corrected by making the DCP semi-automatic instead of fully automatic.

5.4.1 Representative Word a Plural Noun

The reverse inflection algorithm, R_s , transforms text nouns to the nominative singular and attempts to find a match in the dictionary set of representative words. Now if a noun in the set of representative words is in the nominative plural, e.g.,

ЛУЧИ in **КОСМИЧЕСКИЕ ЛУЧИ**/COSMIC RAYS

a member of the noun's paradigm occurring in text will be transformed to the singular form which does not occur in the dictionary, hence no match can be made.

This error can be corrected by altering the reverse inflection algorithm so that it forms, for nouns, both the nominative singular and nominative plural forms.

Unfortunately, such a change will also increase the processing time. It should be pointed out that the great majority of nouns in a piece of text will not be elements of the paradigm of any representative word. Nevertheless, each noun must be processed by R_s to establish that it is or is not a member of a representative word paradigm.

Changing the algorithm R_g so that it forms the nominative plural would require the reprocessing of a given word in its plural forms, if processing of the singular forms fails.

5.4.2 Representative Word a Noun in Adjectival Form

Some Russian nouns are, morphologically, adjectives

(СТОЛОВАЯ/LIVING ROOM , ДАННЫЕ /DATA)

If such a form is a representative word, it will have its adjectival ending removed (subsection 5.2.3). A member of its paradigm occurring in text will have its ending removed. A match will be made, and the English translation will be stored in the Index Matrix. But it will have the wrong part-of-speech – adjective instead of noun. The English translation will thus appear correctly in the simple index, but quite likely will appear incorrectly (if it appears at all) in the complex index. (The rules for forming the complex index could undoubtedly be altered to satisfactorily handle such nouns. This report, however, describes the system as it now stands.)

5.4.3 Miscellaneous Forms

Occasionally, though not often, miscellaneous forms occur in the dictionary such as a preposition followed by a noun, adjective followed by prepositional phrase, or a verb preceded by an adverb. Their occurrence is sufficiently rare that as yet no special provision has been made for handling them by DCP.

5.5 SEMI-AUTOMATIC DICTIONARY

The errors described in the Indexing system can all be traced to the desire to make the Dictionary Creation Program fully automatic. Relaxing this requirement, allowing the DCP to work in tandem with someone possessing a small knowledge of Russian, would, we believe, eliminate the errors discussed.

DCP would, in this version, be a two-pass system. The first pass would create the type of dictionary that has been described. Suspicious entries – those, for example, whose representative words are potential plural nouns – would be printed out with their English translations. A human being would then examine the print-out and make necessary changes. (Thus, a plural noun would be changed by the human to the singular form so that R_s would successfully operate upon a member of its paradigm occurring in text.)

The second pass would simply merge the corrected entries with the dictionary created on the first pass.

5.6 EXAMPLE OF ENTRY SELECTION FOR THE RAW INDEX

... β ЧАСТИЦ ВБЛИЗИ РАДИОАКТИВНЫХ ИСТОЧНИКОВ ПРЕДПОЛОЖИЛ....

...i-4 i-3 i-2 i-1 i i+1 ....

{ИСТОЧНИК , i}

{i , noun , (1,0) , ИСТОЧНИК}

{i , noun , ИСТОЧНИК , РАДИОАКТИВНЫХ ИСТОЧНИКОВ , ИСТОЧНИКОВ}

{i , noun , РАДИОАКТИВН ИСТОЧНИК , ИСТОЧНИК}

{i , noun , 773141651767 , 632463460445}

{i , noun , Radioactive source , i-1 , i}

FILE A

ИСТОЧНИК

noun, (1,0)

FILE B

632463460445

Source

773141651767

Radioactive Source

Finally

F(i-1 , i) (Radioactive source , noun)

Section 6
FINAL INDEX

3.1 GENERAL DISCUSSION

After preliminary processing, the Indexer is ready to construct the index items from the Raw Index. The Raw Index can be viewed as a matrix, M . M , at this point, is a blend of Russian high-frequency words, English phrases constructed from certain Russian words in the original sentence, and the parts of speech of those Russian words. The positions of the Russian words and of the English phrases with respect to each other in the original sentence are preserved in M .

The Indexer can produce two types of Final Index: a Simple Index and a Complex Index.

6.2 SIMPLE INDEX

This is a simple listing, alphabetically arranged and with duplicate entries eliminated, of the English phrases in the raw index and the page numbers on which they occur. There is no cross indexing.

6.3 COMPLEX INDEX

The Complex Index is also formed from the Raw Index. The information contained in the Raw Index -- parts of speech of Russian representative words, sentence limits of the original Russian phrase, and English translation of Russian phrase -- allow index entries to be constructed using syntactic information. The Complex Index is also cross-referenced. Index items are selected by examining the first column of the Raw Index Matrix. This column contains part-of-speech information and original sentence limits of Russian phrases.

Index items are constructed using the syntactic building blocks of noun, adjective, preposition, and verb. Index items are defined below in terms of this syntactic information. To avoid cluttering up the notation, the abbreviations used will be taken to mean the English translation of the part of speech indicated. Thus n does not mean a noun, but a particular English phrase behaving as a noun. Similarly

aj = English phrase behaving as an adjective

v = English phrase behaving as a verb

pr = English word behaving as a preposition

but

pr* = a Russian preposition

pr* is a Russian preposition which cannot be translated with a high degree of accuracy. Some prepositions can be translated with a reasonable degree of accuracy. Their translation is denoted by pr .

The following prepositions are being translated.

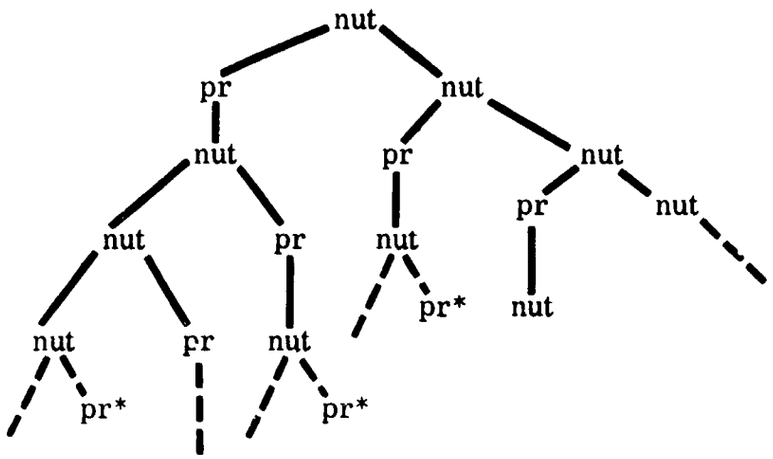
Russian Preposition	Pr
В	in
ДЛЯ	for
ДО	to
К	to
МЕЖДУ	between
О	about
С	with
ПОСЛЕ	after

We also define a noun unit (nut) as follows:

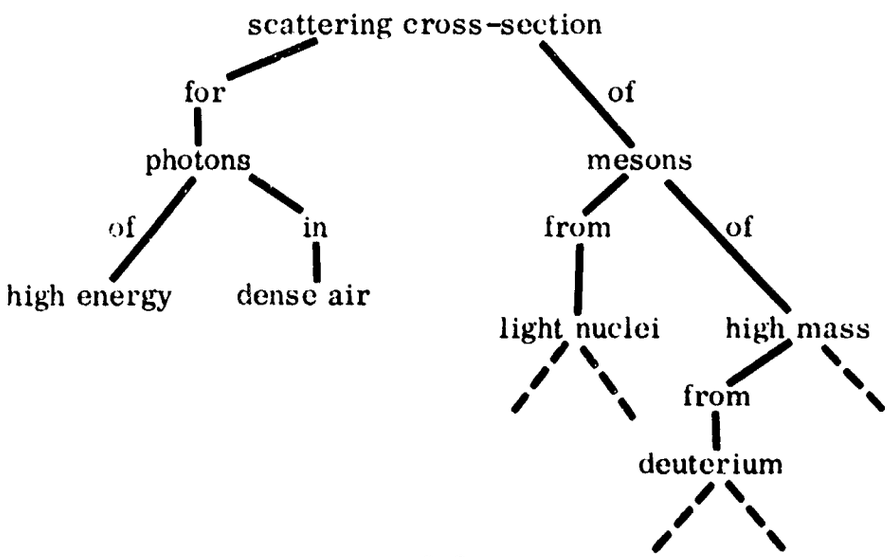
$$\text{nut} \begin{cases} n \\ aj_1 aj_2 \cdots aj_k n \end{cases}$$

In the descriptions to follow it must be remembered that aj , v , n , pr , and nut represent English words and phrases, and pr* represents a Russian word.

The structure of the index items can be shown conveniently by a tree structure. A noun unit appears at the top of the tree. Nodes of the tree below the top node represent English phrases lying to the right of the leading phrase. Adjacent nodes on the tree represent contiguous phrases in the sentence. A branch at a given node indicates that the possibilities indicated may follow the phrase represented by the branch node. A continuous line drawn from the top node through lower nodes gives the structure of an index item except that if pr* occurs in a node the index item terminates at the previous node. If two connected nodes are each a nut, then the English preposition "of" appears on the connecting line to indicate that it is to be inserted between the two noun units.



Example.



Index items for rightmost branch are:

- scattering cross section of mesons of high mass from deuterium
- deuterium, scattering cross section of mesons of high mass from

Index items for sub-branch of rightmost branch are:

- scattering cross section of mesons from light nuclei
- light nuclei, scattering cross section of mesons from

Index items for leftmost branch are:

- scattering cross section for photons of high energy
- high energy, scattering cross section for photons of

Index items for subranch of leftmost branch are:

- scattering cross section for photons in dense air
- dense air, scattering cross section for photons in

The second index item in each case is the cross-indexed entry, the leading noun unit representing the terminal node.

If the leading noun unit contains adjectives as in this example, then deeper cross indexing is possible, leading to the following additional entries:

- cross section, scattering, of mesons of high mass from deuterium
- cross section, scattering, of mesons from light nuclei
- cross section, scattering, for photons of high energy
- cross section, scattering, for photons in dense air

Section 7
CONCLUSIONS

The algorithm which produces English translations for sequences of Russian text words can be viewed as the basic algorithm which, in the system described here, is being used to produce an index.

This basic algorithm has several other possible uses.

7.1 EXTRACTING

Phrases occurring in a technical phrase dictionary are, by definition, descriptors critical to an understanding of a given scientific field. Since it isolates such phrases in text, the basic algorithm can be used to extract Russian sentences, paragraphs, or even pages from a larger body of technical text.

7.2 TRANSLATION

The basic algorithm and the dictionary upon which it operates could be used as a closed subroutine within a larger Russian-English translation system. Such a subroutine would produce translations of a sequence of Russian words which occur in a piece of text. The English translation itself would, of course, have to be inflected to conform to the syntactic use of the phrase within the sentence.

7.3 RETRIEVAL

The basic algorithm gives the capability of creating a unique information retrieval system -- one which accepts English queries and addresses these queries to files of Russian articles, or more accurately, to files of indexes of Russian articles.

The Russian Retrieval Program follows directly in conception from the Russian-English Indexing System described in this report and will, in fact, use most of the computer programs used by the Indexing System.

Retrieval of articles processed by the Indexing System appears to be simple. Russian articles so processed have been deeply indexed in English. If a user seeking information from a file of such deep indexes uses the same terminology (i.e., the manual version of the computer dictionary) as was used to create the index, then a matching process – user's phrases versus index – in combination with the logical AND and OR operations will enable the user to address long English queries to the file. In effect this will lead to retrieval by English queries of Russian technical material.

The Indexing and Retrieval applications have been discussed with the intention of deriving English information from Russian text. The logic involved, however, applies to English as well. Thus the Russian-English programs with minor alterations may be used to index (Ref. 7), extract, and retrieve English technical material.

Section 8
REFERENCES

8.1 CITED REFERENCES

1. Russian-English Geological Dictionary, compiled by T. A. Sofrano, Moscow, Central Editorial Board, Foreign-Language Scientific and Technical Dictionaries, 1960

2. Ludmilla Ignatius Callahan, Russian-English Technical and Chemical Dictionary, New York, Wiley, 1947

3. D. I. Voskoboinik and M. H. Zimmerman, Russian-English Nuclear Dictionary, Moscow, Central Editorial Board, Foreign-Language Scientific and Technical Dictionaries, 1960

4. N. I. Dozorov, English-Russian Radio-Electronics Dictionary, Military Publishing House of the Ministry of Defense of the U.S.S.R., Moscow, 1959

5. L. P. German-Prosorova and N. I. Vinogradova, English-Russian Radio Technical Dictionary, Central Editorial Board, Foreign-Language Scientific and Technical Dictionaries, Moscow, 1960

6. A. M. Murashkevich, English-Russian Rocket Dictionary, State Publishing Office for Physical and Mathematical Literature, Moscow, 1958

7. D. L. Smith and L. Earl, "English Text Indexing System," Information Retrieval Note No. 52, Dec 1964 (Unpublished but available at LMSC)

8.2 UNCITED REFERENCES

Forbes, Neville, Russian Grammar, Oxford University Press, 1915. Contains an excellent detailed description of Russian grammar and the inflection of nouns and adjectives.

Oettinger, A. G., Automatic Language Translation, Harvard University Press, Cambridge, Massachusetts, 1960. Discusses the machine translation problem with respect to Russian, including a discussion of the problems encountered in building a Russian computer dictionary.

Appendix A

THE FIRST AND SECOND VERB CONJUGATIONS

First Conjugation

Imperfective	*Perfective*	
I. Infinitive:		
чита́ть to read, be reading	прочита́ть	to have read
II. Indicative:		
	Present Tense	
I read, am reading		
я чита́ю ты чита́ешь он, она́, оно́ чита́ет		None
мы чита́ем вы чита́ете они́ чита́ют		
	Past Tense	
I read, was reading	I have, had read	
я чита́л, ла, ло ты чита́л, ла, ло он чита́л она́ чита́ла оно́ чита́ло	я прочита́л, ла, ло ты прочита́л, ла, ло он прочита́л она́ прочита́ла оно́ прочита́ло	
мы, вы они́ чита́ли	мы, вы, они́ прочита́ли	
	Future Tense	
I shall read, be reading	I shall have read	
я бу́ду чита́ть ты бу́дешь чита́ть он, она́, оно́ бу́дет чита́ть	я прочита́ю ты прочита́ешь он, она́, оно́ прочита́ет	
мы бу́дем чита́ть вы бу́дете чита́ть они́ бу́дут чита́ть	мы прочита́ем вы прочита́ете они́ прочита́ют	

Imperfective		*Perfective*	

III. Subjunctive (conditional):

Conjugated exactly like the *past* tense of the *indicative* mood with the addition of particles бы or б:

я читáл, ла, ло бы (б) etc.		я прочитáл, ла, ло бы (б) etc.	
I should read, be reading, should have been reading		I should have read	

IV. Imperative:

читáй! читáйте!	read!	прочитáй! прочитáйте!	read! read (it) through, completely

V. Adverbial participles:

Present Tense

читáя	reading, while reading	None	

Past Tense

читáвши читáв	while (I, etc.) was reading	прочитáвши прочитáв	having read

VI. Participles:

a. *Active*:

Present Tense

читáющий	one who is reading	None	

Past Tense

читáвший	one who was reading	прочитáвший	one who has, had read

Imperfective	*Perfective*	

b. *Passive*:

Present Tense

Long form: чита́емый Short form: чита́ем	None	

which is being read

Past Tense

Long form: чи́танный	прочи́танный	which has, had
Short form: чи́тан	прочи́тан	been read

which was read

(Other *past passive participle* endings are: long -тый, short -т.)

VII. **Passive:**

The *passive* is constructed by means of the *short passive participle* forms, *present* or *past* (see directly above); also by means of the *reflexive* form.

Second Conjugation

Imperfective	*Perfective*	
I. Infinitive:		
кури́ть to smoke, be smoking	**вы́курить**	to have smoked
II. Indicative:		
	Present Tense	
I smoke, am smoking		None
я курю́ **ты ку́ришь** он, она́, оно́ **ку́рит**		
мы ку́рим **вы ку́рите** они́ **ку́рят**		
	Past Tense	
I smoked, was smoking		I have, had smoked
я кури́л, ла, ло **ты кури́л, ла, ло** он кури́л она́ кури́ла оно́ кури́ло		**я вы́курил, ла, ло** **ты вы́курил, ла, ло** он вы́курил она́ вы́курила оно́ вы́курило
мы, вы, они́ кури́ли		мы, вы, они́ вы́курили
	Future Tense	
I shall smoke, be smoking		I shall have smoked
я бу́ду кури́ть **ты бу́дешь** кури́ть он, она́, оно́ **бу́дет** кури́ть **мы бу́дем** кури́ть **вы бу́дете** кури́ть они́ **бу́дут** кури́ть		**я вы́курю** **ты вы́куришь** он, она́, оно́ **вы́курит** мы вы́курим вы вы́курите они́ вы́курят
III. Subjunctive (conditional):		
Conjugated exactly like the *past* tense of the *indicative* mood with the addition of particles **бы (б)**:		
я кури́л, ла, ло бы (б) etc.		**я вы́курил, ла, ло бы (б)** etc
I should smoke, be smoking, should have been smoking		I should have smoked

Imperfective		*Perfective*	
IV. Imperative:			
курй! курйте!	smoke!	вы́кури! вы́курите!	smoke! finish smoking!
V. Adverbial participles:			
	Present Tense		
куря́	smoking, while smoking		None
	Past Tense		
курйвши курйв	while (I, etc.) was smoking	вы́куривши вы́курив	having smoked
VI. Participles:			
a. *Active*:			
	Present Tense		
куря́щий	one who is smoking		None
	Past Tense		
курйвший	one who was smoking	вы́куривший	one who has, had smoked
b. *Passive*:			
	Present Tense		
Long form: курймый Short form: курйм			None
which is being smoked			
	Past Tense		
Long form: ку́реный Short form: ку́рен		вы́куреный вы́курен	which has, had been smoked
which was smoked			

(Other *past passive participle* endings are long **-тый**, short **-т**.)

VII. Passive:

The *passive* is constructed by means of the *short passive participle* forms, *present* or *past* (see directly above); also by means of the *reflexive* form.

Appendix B
REVERSE INFLECTION ALGORITHM

The algorithm operates on a dictionary which contains classically defined canonical forms: nominative singular of nouns, nominative singular masculine gender for adjectives, and the infinitive for verbs and participles.

A word encountered in text has a terminal string of letters removed by the algorithm and a new terminal string added to form a "pseudo-word." The pseudo-word is an attempt on the part of the algorithm to construct the text word's canonical form. If the pseudo-word does in fact exist in the dictionary, the algorithm proceeds to examine the next word.

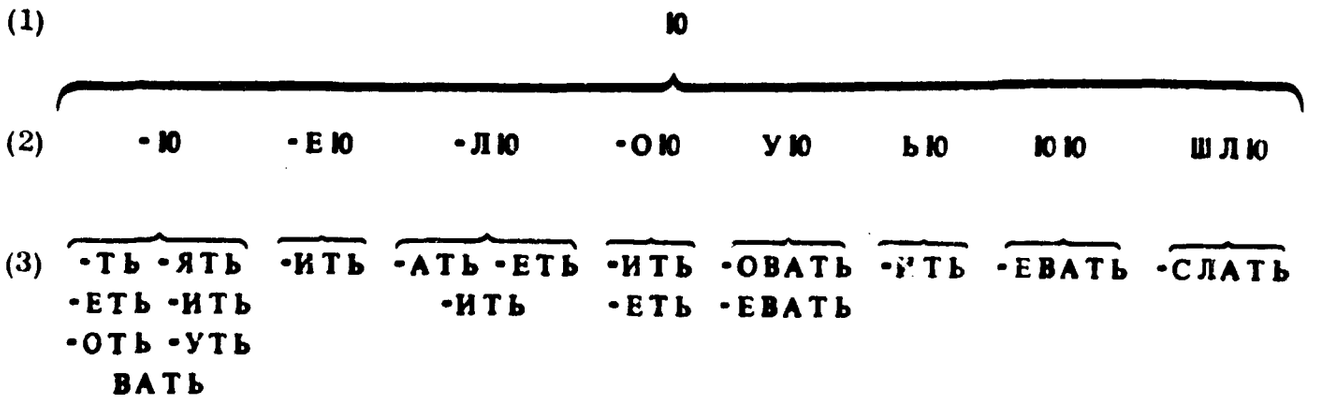
If the pseudo-word does not exist in the dictionary, a new pseudo-word is constructed (terminal string removed, new terminal string added). The process continues for a given word until a true canonical form is constructed or until all of the text word's possible constructions have been exhausted.

Pseudo-word construction takes place for all of a word's potential parts-of-speech. Thus, the algorithm assumes a word is a verb, noun, adjective, participle, in that order, and constructs, if possible, a set of pseudo-words for each part-of-speech.

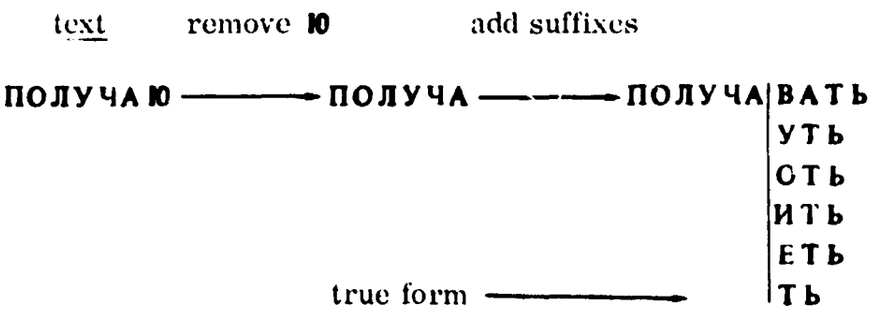
The algorithm operates upon a table which has a 3-level structure. The levels are as follows:

(1) Terminal letter for a given part of speech
(2) Possible suffixes ending in the terminal letter for this part of speech
(3) Canonical suffixes to be added to the stem after suffixes of level 2 have been removed

Example: This example shows the 3-level structure for verbs ending in Ю.



Now suppose ПОЛУЧАЮ occurs in text. The algorithm assumes first it is a verb. It examines the terminal letter, finds that it is Ю and that Ю has eight possible verbal suffixes. In this case, only a single suffix, Ю, is contained in the word. Ю is removed from the text word and the seven canonical suffixes are added to the stem to form seven pseudo-words. The suffix (-ТЬ) gives a true canonical form. If, now (ПОЛУЧАТЬ) is in the dictionary, a match will be made.



It may be objected that the algorithm simulates too closely human processes, that it is illogical - inelegant, even - to remove strings of letters only to add new strings. Why not, for example, restrict the algorithm to the removal operation? That is, make the canonical entries in the dictionary be some stripped form of the real word. Then the algorithm need only remove endings and compare the stripped word to canonical entries which have also been stripped.

At the time the algorithm was developed, it was considered desirable to put as much of the translation burden as possible on the computer. If a stripped form of a word is used as the canonical entry, how is the stripped form arrived at? It must be decided by human analysis or by some computer algorithm. Either way, additional labor, human or machine, is necessary. It is not, then, a question of two operations versus one – removal and addition versus removal – but of two operations versus two operations – removal and addition versus removal and removal. In the first case, removal and addition occur in the same algorithm. In the second case, the first removal operation is performed by a human or a computer, but in either case it takes place independently and prior to the second removal operation.

Further, the algorithm was designed for use with a parsing program and it was felt that there was important syntactic information which was characteristic of classically defined words that would disappear if a true word were reduced, in effect, to a "non-word." (We are thinking here of the phenomena of syntactic and semantic government.)

Appendix C
COMPRESSION OF PHRASES INTO A LOGICAL SUM

The phrase to be compressed is

РАССТОЯНИЕ МЕЖДУ ЧАСТИЦАМИ
interparticle distance

The IBM 7094 allows six alphanumeric characters to be stored in a single machine word.

Р	А	С	С	Т	О
Я	Н	И	Е	0	М
Е	Ж	Д	У	0	Ч
А	С	Т	И	Ц	А
М	И	0	0	0	0

Note that blanks have been replaced by zeros and that the final machine word has also been padded out with zeros.

The logical sum can now be calculated. (The actual numbers shown below are the numerical representation in core of the corresponding Russian letters above.)

	51	21	62	62	63	46
	13	45	31	25	00	41
	25	53	24	64	00	04
	21	62	63	31	23	21
	44	31	00	00	00	00
	00	56	24	25	07	37
						2
Final Sum	00	56	24	25	07	41

Appendix D

CONCEPTUAL DESCRIPTION OF RUSSIAN TEXT PHRASE TRANSLATOR

D. 1 INTRODUCTION

The Russian text phrase translator consists of two operationally distinct parts: the first being a computer-generated phrase dictionary, the second a computer-generated textual analysis which assigns to every text interval an English phrase and grammatical function. The English phrase is either a translation of the text interval or a statement to the effect that the text interval is neither a dictionary phrase nor an inflected form of a dictionary phrase.

In this description of the Translator, the following notations will be adhered to. Collections of phrases will be denoted by capital script letters: $\mathcal{R}$, $\mathcal{E}$, ... $\mathcal{R}_a$, $\mathcal{R}_b$, ..., will denote the subfamily of $\mathcal{R}$ all of whose translations begin with the letter a , b , ... Phrases will be denoted by capital Roman letters: R , E , ... The English translation of a Russian phrase R will be denoted by E(R) . Words will be denoted by small Greek letters: ω , μ , ... Small Roman letters will denote numbers or themselves. To any Russian word θ [phrase R] is associated its canonical form θ' [R] and its numerical logical sum $g(\theta)$. [g(R)] .

D. 2 THE PHRASE DICTIONARY

Let $\mathcal{R}$ be a family of Russian phrases in which the left-most noun or verb occurs in canonical form. Let $\mathcal{E}(\mathcal{R})$ be the family of English translations. For each $R \in \mathcal{R}$, a representative word ω in canonical form is algorithmically determined ($\omega \in R$) as well as its part of speech (POS) and its imbedding coordinates $x(\omega, R)$, $y(\omega, R)$ relative to R , that is, if $R = \theta_1\theta_2 \ldots \theta_i(=\omega)\,\theta_{i+1} \ldots \theta_{i+j}$ then $x(\omega, R) = i - 1$, $y(\omega, R) = j$. Let now $\mathcal{R}_a(\omega) \subset \mathcal{R}_a$ be that subfamily of $\mathcal{R}_a$ whose canonical representative word is ω . The a-maximal coordinates of ω are defined as

$$x(a) = \max_{R \in \mathcal{R}_a(\bar{\omega})} x(\bar{\omega}, R) \qquad y(a) = \max_{R \in \mathcal{R}_a(\bar{\omega})} y(\bar{\omega}, R)$$

Conceptually, a typical entry in the phrase dictionary is

$$\left(g(R), E(R), \bar{\omega}, POS, x(\cdot), y(\cdot)\right)$$

The a-section of the dictionary may be written as

$$\bigcup_{R \in \mathcal{R}_a} \left(g(R), E(R), \bar{\omega}, POS, x(a), y(a)\right)$$

and the entire phrase dictionary is

$$\bigcup_{(\cdot)=a}^{z} \bigcup_{R \in \mathcal{R}_{(\cdot)}} \left(g(R), E(R), \bar{\omega}, POS, x(\cdot), y(\cdot)\right)$$

D.3 THE TEXT ANALYSIS

A prescribed Russian text will be regarded as an ordered set of n words. Each and every word interval (a, b) $a \leq b$ will be regarded as a text phrase. Denote by $\mathcal{T}$ the set of text phrases, by $\mathcal{E}(\mathcal{R})$ the set of English phrases in the phrase dictionary. Conceptually the textual analysis may be signified as

$$F: \mathcal{T} \rightarrow \left(\mathcal{E}(\mathcal{R}), \text{grammatical function}\right)$$

Operationally only the dictionary significant phrases in $\mathcal{T}$ are analyzed, the others being assigned an English phrase by fiat.

The analysis, interpretation, and production of data by the textual analysis algorithm are perhaps most succinctly described in the following conceptual flow chart. The notations are as described in the introduction. Let $\mathcal{D}_a \equiv \bigcup_{R \in \mathcal{R}_a} \left(g(R), E(R), \bar{\omega}, POS, x(a), y(a)\right)$ be the a-section of the phrase dictionary. Let ω_i be the i-th word in the ordered text:

STEP 1: Canonical form

$$(\omega, i) \rightarrow (\bar{\omega}, i)$$

STEP 2: Look up $\bar{\omega}$ in $\mathcal{D}_a$ (if not successful go to $\mathcal{D}_b, \ldots \mathcal{D}_z$. Failing, go to ω_{i+1})

$$(\bar{\omega}, i) \rightarrow \left(\bar{\omega}, i, POS, x(a), y(a)\right)$$

STEP 3: Formation of phrases in which ω is imbedded

$$\left(\bar{\omega}, i, POS, x(a), y(a)\right) \rightarrow \{(\bar{\omega}, i, POS, T)\}\ ,\ \omega \in T \subset \left(i - x(a), i + y(a)\right)$$

STEP 4: Canonical form

$$\{(\bar{\omega}, i, POS, T)\} \rightarrow \{(i, POS, \bar{T})\}$$

STEP 5: Logical sum

$$\{(i, POS, \bar{T})\} \rightarrow \left\{\left(i, POS, g(\bar{T})\right)\right\}$$

STEP 6: Phrase dictionary look up over $\{g(\bar{T})\}$

The set of $x(a)$ - $y(a)$ triples $\left(i, POS, g(\bar{T})\right)$ are ordered according to the length of the phrase T, the ordering amongs equal length phrases being indifferent. The ordered logical sums $\{g(T)\}$ are then matched against the set $\{g(\bar{R})\}$ appearing in $\mathcal{D}_a$, etc. Assuming the first agreement occurs for $g(\bar{S}) \in \mathcal{D}_{(\ell)}$, the algorithm then produces the information vector

$$\left\{\left(i, POS, g(T)\right)\right\} \rightarrow \left(i, POS, g(S), E(S), x(\alpha, S), y(\alpha, S)\right)$$

STEP 7

$$F\left(i - x(\alpha, S), i - y(\alpha, S)\right) \rightarrow \left(F(S), POS\right)$$

STEP 8: After an agreement has occurred [if no agreement occurs in STEP 6], the textual analysis algorithm is repeated for $\omega_{i+y(\bar{\omega},S)+1}$ $[\omega_{i+1}]$ provided the indicated subscript is $\leq n$. Otherwise the text phrase translation is terminated by assigning to all non determined text interval phrases $T \in \mathcal{T}$ the English phrase "not significant," and grammatical function "none."

Appendix E
OUTPUT OF THE INDEXING SYSTEM

The first few paragraphs of a geological article entitled "Phase Transformations in the Interior of the Earth," by S. M. Stishov, Nature, Sep 1962 were used as input to the Indexer (Fig. E-1). The simple and complex indexes are listed in Figs. E-2 and E-3. Single words have been eliminated from the complex index, though not from the simple. The computer dictionary is based on Sofiano's geological dictionary (Ref. 1).

Гипотезы

ФАЗОВЫЕ ПРЕВРАЩЕНИЯ В ГЛУБИНАХ ЗЕМЛИ

Достижения механики и физики в XVII—XVIII вв. позволили определить массу и среднюю плотность Земли. Последняя оказалась равной 5,5 *г/см*³. А так как плотность наиболее тяжелых пород на поверхности Земли не превышает 3,3 *г/см*³, то, естественно, возникло представление, что плотность Земли увеличивается с глубиной.

Факты существования железных метеоритов, а также в прошлом популярная теория происхождения Земли из горячего вещества Солнца привели многих ученых к мысли о концентрации железа в центре Земли. Примечательно, что уже вполне определенные высказывания французского геолога А. Добрэ в 1866 г. о железном ядре Земли вскоре получили поддержку со стороны сейсмологов, которым в конце XIX и начале XX в. удалось установить наличие в Земле ядра.

В 20-х годах текущего столетия В. М. Гольдшмидт (Норвегия) и немецкий физикохимик Г. Тамман развили представление о том, что в первоначально расплавленной Земле происходило разделение (дифференциация) веществ по их плотности, аналогично тому, что мы имеем, например, при плавке сульфидных руд. При этом процессе появляются три слоя: шлак (силикатный слой), штейн (смесь сульфидов и металла) и собственно металл. Согласно этой гипотезе, в Земле выделялись следующие слои: силикатный и сульфидный (оболочка Земли) и металлический, состоящий из железа с примесью никеля (ядро Земли).

Американские ученые Ф. Кларк, Г. Вашингтон, Л. Адамс и др. не выделили сульфидный слой; они полагали, что между железным ядром и силикатной оболочкой находится промежуточная область, состоящая из смеси силикатов и железа.

Теория слоистой, химически дифференцированной Земли, во многом подкреплялась данными сейсмологов, которые первоначально считали, что в мантии (оболочке) Земли, т. е. в той ее части, которая расположена между земной корой и ядром, существует много границ раздела.

В дальнейшем успехи геофизики и космогонии, связанные главным образом с имена-

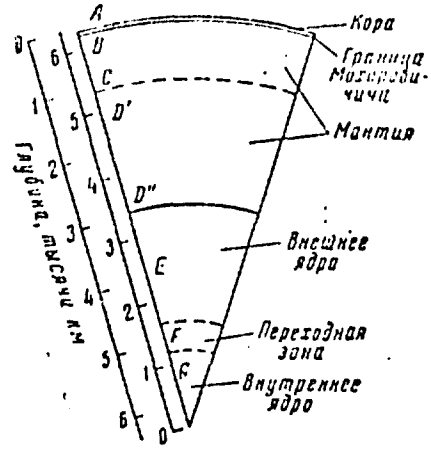


Рис. 1. Зоны в Земле по *К. Буллену*

Fig. E-1 Russian Text Used for English Indexing

PAGE 105

AVERAGE DENSITY
CENTER
CHEMICAL
COATING
DENSITY
DIFFERENTIATED
DIFFERENTIATION
EARTH
EARTH SHELL
GEOLOGIST
HYPOTHESIS
IMPURITY
INTERMEDIATE
IRON
IRON METEORITE
LIMIT
MANTLE
MATTER
MELTING
METAL
METALLIC
MIXTURE
ORIGIN
PROCESS
REGION
ROCK
SIDE
SILICATE
SLAG
STRATUM
SULPHIDE ORE
SULPHITE
SUPPORT
SURFACE
THEORY

Fig. E-2 Simple Index

AVERAGE DENSITY OF EARTH
COATING,SILI ATE,
DENSITY OF EARTH
DIFFERENTIATION OF MATTER
EARTH, AVERAGE DENSITY OF
EARTH, DENSITY OF
EARTH, IRON IN CENTER OF
EARTH, SURFACE OF
EARTH, THEORY OF ORIGIN OF
EARTH SHELL
EXISTENCE OF IRON METEORITE
IMPURITY, IRON WITH
INTERMEDIATE REGION
IRON IN CENTER OF EARTH
IRON WITH IMPURITY
IRON METEORITE, EXISTENCE OF
MATTER, DIFFERENTIATION OF
MELTING OF SULPHIDE ORE
METAL, MIXTURE OF SULPHITE AND
MIXTURE OF SULPHITE AND METAL
REGION,INTERMEDIATE,
SILICATE COATING
SILICATE STRATUM
STRATUM,SILI ATE,
SULPHIDE ORE, MELTING OF
SURFACE OF EARTH
THEORY OF ORIGIN OF EARTH

Fig. E-3 Complex Index

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

National Security Agency
Fort George G. Meade, Maryland
Attn: Librarian, C-332

Lincoln Laboratory
Massachusetts Institute of Technology
Lexington 73, Massachusetts
Attn: Library

Daniel E. Kaplan
Dept. 52-40, Bldg. 202
Lockheed Missiles & Space Company
3251 Hanover Street
Palo Alto, California

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author) Lockheed Palo Alto Research Laboratory Lockheed Missiles & Space Company Palo Alto, California	2a REPORT SECURITY CLASSIFICATION Unclassified 2b GROUP N/A

3 REPORT TITLE

Annual Report: Automatic Indexing and Abstracting, Part II. English Indexing of Russian Technical Text

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)

Annual Progress Report

5 AUTHOR(S) (Last name, first name, initial)

Rudin, B. D., principal investigator
Robison, H. R.

6 REPORT DATE	7a TOTAL NO. OF PAGES	7b NO. OF REFS
March 1966	57 text pages	7 cited, 2 uncited

8a CONTRACT OR GRANT NO.	9a ORIGINATOR'S REPORT NUMBER(S)
Nonr 4440(00) b. PROJECT NO. c. d.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) M-21-66-2

10 AVAILABILITY/LIMITATION NOTICES

The complete report is available in the major Navy technical libraries and can be obtained from the Defense Documentation Center. A few copies are available for distribution by the author.

11 SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY

13 ABSTRACT

The following report describes a computer system for the IBM 7094 which produces English indexes of technical Russian text.

Part of the indexing system produces a machine dictionary on magnetic tape. This dictionary is a computer representation of standard English-Russian technical phrase dictionaries.

The indexing portion of the system matches Russian text phrases against Russian dictionary phrases. Dictionary phrases are in canonical form; reverse inflection algorithms transform text phrases to their canonical form. When a match is found, the English translation of the match is extracted from the dictionary. The final index is constructed from set of such English translations.

DD FORM 1 JAN 64 1473

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.