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

ED 012 017 AL 000 435

ANNOTATED BIBLIOGRAPHY OF RAND PUBLICATIONS IN COMPUTATIONAL LINGUISTICS.

BY- HAYS, DAVID G. AND OTHERS
RAND CORP., SANTA MONICA, CALIF.

REPORT NUMBER RM-3894-3 PUB DATE JUN 67
EDRS PRICE MF-$0.09 HC-$1.88 47P.

DESCRIPTORS- *COMPUTATIONAL LINGUISTICS, *LINGUISTIC THEORY, *PSYCHOLINGUISTICS, RUSSIAN, *INFORMATION RETRIEVAL, MACHINE TRANSLATION, MATHEMATICAL LINGUISTICS, CYBERNETICS, DOCUMENTATION, LANGUAGE TYPOLOGY, SANTA MONICA

THIS REVISED ANNOTATED BIBLIOGRAPHY LISTS 143 RAND PUBLICATIONS IN COMPUTATIONAL LINGUISTICS, INCLUDING SUCH AREAS AS LINGUISTIC RESEARCH METHODS, STUDIES ON THE RUSSIAN AND ENGLISH LANGUAGES, INFORMATION RETRIEVAL, PSYCHOLINGUISTICS, AND CHARACTER READERS. ENTRIES ON THE RUSSIAN LANGUAGE ARE FURTHER ORGANIZED AS ANALYSES OF TEXTS AND GLOSSARIES, SYNTAX, AND SEMANTICS, WHILE ANOTHER SECTION OF THE BIBLIOGRAPHY COVERS PROBLEMS IN MACHINE TRANSLATION. THIS PUBLICATION ALSO INCLUDES A LIST OF FOREIGN AND DOMESTIC DEPOSIT LIBRARIES WHERE COPIES OF ALL REPORTS CONTAINED IN THE BIBLIOGRAPHY CAN BE FOUND. WITH THE EXCEPTION OF COMMERCIALLY PUBLISHED BOOKS BY RAND AUTHORS AND ARTICLES THAT HAVE APPEARED IN COMMERCIAL EDITIONS, ALL ENTRIES AND THE BIBLIOGRAPHY ITSELF ARE AVAILABLE DIRECTLY FROM THE RAND REPORTS DEPARTMENT, 1700 MAIN ST., SANTA MONICA, CALIFORNIA 90406. (FB)

MEMORANDUM

RM-3894-3

JUNE 1967

ANNOTATED BIBLIOGRAPHY OF RAND PUBLICATIONS IN COMPUTATIONAL LINGUISTICS

David G. Hays, Bozena Henisz-Dostert and Marjorie L. Rapp

The RAND Corporation
1700 MAIN ST. • SANTA MONICA • CALIFORNIA • 90406

PREFACE

This bibliography lists and annotates 143 RAND publications on linguistic theory, linguistic research methods, computational methods for linguistics, the Russian language, the English language, information retrieval, automatic content analysis, psycholinguistics, and character readers. In order to keep the size and scope of the bibliography within convenient limits, other studies on such subjects as information theory (in the ordinary sense of probabilistic analysis of communication channels) and artificial intelligence in general (including nerve-net simulation) have been omitted here. References to each related RAND publication can, however, be found in the one-volume Index of Selected Publications of The RAND Corporation, 1946-1962, and in its successor, Selected RAND Abstracts, a quarterly series collected in annual volumes beginning in 1963.

The new or completely revised entries in this third revision of the Bibliography are 1-21, 1-22, 1-23, 1-24, 1-25, 2.2-2, 2.2-6, 2.2-7, 2.2-8, 2.2-9, 3-9, 3-10, 4.1-6, 4.2-16, 4.2-17, 5.1-1, 5.1-5, 5.2-8, 5.2-9, 5.2-10, 6-3, 7-13, 7-14, 7-15, 7-16, 7-17, 8-3, 9-11, 9-12, and 10-6.

Aside from studies that have appeared in commercial editions, the publications listed in this bibliography may be obtained by writing directly to the RAND Reports Department (1700 Main Street, Santa Monica, California 90406). Wherever a DDC document (AD) number is given, they are also available

from the Defense Document Center (Cameron Station, Alexandria, Virginia 22314), or from the Clearinghouse for Federal Scientific and Technical Information (U.S. Department of Commerce, Springfield, Virginia 22151). From the latter two sources, requests will be filled most quickly if the AD numbers provided in the citations are used.

Virtually all of the reports have been deposited with the libraries indicated on pp. 24-27. There they are available to library patrons on the same basis as other holdings. They may also be borrowed through the Interlibrary Loan Service.

Commercially published books by RAND authors are available only through the publisher or a bookseller. Aside from such studies, the publications listed in this bibliography may be purchased by the public directly from the RAND Reports Department. The price schedule listed below was established to cover costs of reproduction and handling.

Page Count*	Price**
1-49	$1.00
50-99	2.00
100-199	3.00
200 and over	4.00

---

* Page counts are included in each publication citation.

** California residents add 4 percent sales tax.

CONTENTS

Annotated Bibliography of RAND Publications in Computational Linguistics

1. General

1. NEW AREAS OF APPLICATION OF COMPUTERS
A. Newell
P-2142, November 1960, 4 pp.
Datamation, vol. 7, no. 1 (January 1961), pp. 15-16.

In 1960, computers were in use for engineering calculation, business-data processing, preparation of Bible concordances, prediction of election returns, discovery of new chemical names, composition of elementary music, solution of double-crostics, and playing solitaire. Future applications suggested were the handling of large hydrodynamic problems and of raw language, and the development of elementary language-understanding programs.

2. BENCHMARKS IN ARTIFICIAL INTELLIGENCE
F. J. Gruenberger
P-2586, June 1962, 13 pp.
Datamation, vol. 8, no. 10 (October 1962), pp. 33-35.

A list of tasks, presently performable by most human beings, that may eventually be handled successfully by computers. These tasks can serve as milestones in the field of artificial intelligence.

3. SOVIET CYBERNETICS AND COMPUTER SCIENCES, 1960
E. A. Feigenbaum
RM-2799-PR, October 1961, 77 pp.
AD 266 129
IRE Transactions on Electronic Computers, vol. EC-10, no. 4 (December 1961), pp. 759-776.

Feigenbaum describes his experiences as a delegate to the International Congress on Automatic Control held at Moscow, June 27 - July 7, 1960. He discusses (i) certain aspects of the conference; (ii) some Soviet research projects in artificial intelligence and biocybernetics; and (iii) general Soviet attitudes, techniques, and directions in the cybernetic and computer-related sciences. He concludes that Soviet research in the computer sciences lags behind Western development, but that the gap is neither large nor based on a lack of understanding of fundamental principles. The Soviets will progress rapidly if and when priority, in terms of access to computing machines, is given to their research.

4. AUTOMATIC LANGUAGE-DATA PROCESSING
D. G. Hays
(Unsponsored)
In: Harold Borko, ed., Computer Applications in the Behavioral Sciences, Prentice-Hall, Englewood Cliffs, N. J., 1962, pp. 394-421.

ALDP includes content analysis (i.e., of novels, editorials, discussion-group transcripts, interviews), MT, etc. Translation is taken as illustrative: Input and output problems, dictionary lookup, sentence-structure determination, and output construction are discussed. Phrase-structure and dependency theories are briefly compared. The first routine capable of finding all structures for a given sentence under a given grammar with a feasible number of operations (due to John Cocke) is reported, and compared with the earlier RAND program that could find at most one structure per sentence. Semantics, and its relation to information-retrieval and psychological ALDP systems, is treated briefly.

5. STUDIES IN MACHINE TRANSLATION--1: SURVEY AND CRITIQUE
H. P. Edmundson, K. E. Harper, and D. G. Hays
RM-2063, February 1958, 28 pp.
AD 150 672

This very early survey noted that both words and sentence structures had to be translated. Glossary compilation and lookup, idioglossaries, thesaurus methods, idiom recognition, translation of prepositions, construction of grammars by semiautomatic methods, early sentence-structure determination routines, and hardware requirements for input, storage, and output were reviewed. Great stress was laid on empirical linguistics.

6. THE USE OF MACHINES IN THE CONSTRUCTION OF A GRAMMAR AND COMPUTER PROGRAM FOR STRUCTURAL ANALYSIS
K. E. Harper and D. G. Hays
P-1588, January 1959, 15 pp.
In: Information Processing, UNESCO, Paris, 1960, pp. 188-194.
Also (in Russian) In: L. N. Zasorina, ed., Avtomatizatsiya v Lingvistike, Izdatel'stvo "Nauka", Moscow, 1966, pp. 132-143 (S. Ya. Fitialova, tr.).

The cyclic research process is briefly described, including mention of the use of structural concordances in identification of new syntactic classes. The RAND sentence-structure determination routine, based on precedence and dependency theory, is outlined.

7. LINGUISTIC RESEARCH AT THE RAND CORPORATION
D. G. Hays
P-1900, February 1960, 22 pp.
In: H. P. Edmundson, ed., Proceedings of the National Symposium on Machine Translation, Prentice-Hall, Englewood Cliffs, N. J., 1961, pp. 13-25.

In 1960, the RAND project was interested in postediting as a basis for linguistic research, the use of semi-automatic research methods to follow postediting, and distributional analysis of transformations and semantics. The notion of idiom-in-structure is mentioned here: it is a fixed combination of words having distinctive meaning, but free to occur in different orders.

8. LINGUISTIC ANALYSIS IN MACHINE TRANSLATION RESEARCH
H. P. Edmundson
P-1328, April 1958, 10 pp.
In: Martha Boaz, ed., Modern Trends in Documentation, Pergamon Press, New York, 1959, pp. 31-37.

A popularized exposition of some fundamental ideas in linguistics and mathematics.

9. AUTOMATIC COMPUTERS IN MACHINE-TRANSLATION RESEARCH
D. G. Hays
P-1321, April 1958, 10 pp.
In: Martha Boaz, ed., Modern Trends in Documentation, Pergamon Press, New York, 1959, pp. 38-44.

A popularized exposition of the cyclic research method, of phrase-structure parsing, and of requirements for computers with special features.

10. SIX TASKS IN COMPUTATIONAL LINGUISTICS
K. E. Harper, D. G. Hays, D. S. Worth, and T. W. Ziehe
RM-2803-AFOSR, October 1961, 114 pp.
AD 264 769
(Air Force Office of Scientific Research: AFOSR-TN-1362)

(i) Distributional semantics. 42 distributional classes are listed. For example, some verbs always take an animate subject (103 such verbs were found in a block of about 100,000 occurrences of physics text).
(ii) Discovery procedures for transformations. If two structures are transformationally linked in the sence of Harris, they should tend to occur in text with about the same words filling them.
(iii) Development of a Russian word-family glossary. The forms in the RAND physics glossary were segmented by synchronic criteria into constituent morphs. About 2,500 families, each having a common root, were identified.
(iv) A text processor. This system is used to transcribe printed pages of text on tape, list from tape, and make corrections. Codes and formats are given.
(v) A glossary editor. Preliminary discussion of a system for maintenance of a deeply organized glossary file.
(vi) Text-improvement program. Discussion of errors in the RAND text file, including both raw text and editors' notes, and of procedures applied to eliminate them.

11. DICTIONARY PROBLEMS IN MACHINE TRANSLATION
K. E. Harper
P-2327, May 1961, 15 pp.
In: Paul L. Garvin, ed., Natural Language and the Computer, McGraw-Hill Book Co., New York, 1963, pp. 215-222.

Punched-card and magnetic disk or drum systems are inherently slow. Tape systems have been used in several ways: (i) Alphabetized text is compared with alphabetized dictionary. (ii) Text is compressed into core storage and the dictionary read from tape. (iii) The dictionary is compressed into storage by use of letter-table addressing. The second and third schemes are fast enough to make the need for special-purpose equipment doubtful. Treatment of "idioms" (input units containing internal blanks) and of word segmentation are mentioned, and methods of collection of entries for MT dictionaries are considered. Harper also discusses the form and content of grammar and semantic codes and target-language equivalents.

12. SOVIET RESEARCH IN MACHINE TRANSLATION
K. E. Harper
P-1896, February 1960, 19 pp.
AD 244 725
In: H. P. Edmundson, ed., Proceedings of the National Symposium on Machine Translation, Prentice-Hall, Englewood Cliffs, N. J., 1961, pp. 2-12.

In 1960, a survey of the available literature indicated that Soviet researchers were working hard on MT, developing glossaries and routines for input analysis, transfer, and output synthesis, but also developing sophisticated linguistic theories applicable to the problem. Computing facilities seemed to be unavailable except for short demonstrations.

13. ON CYBERNETICS, INFORMATION PROCESSING, AND THINKING
M. E. Maron
P-2879, March 1964, 41 pp.
AD 435 484
In: N. Wiener and J. P. Schade, eds., Cybernetics of the Nervous System (Progress in Brain Research, vol. 17), Elsevier, Amsterdam, 1965, pp. 118-138.

It is the purpose of this paper to examine the origins, development, and present status of those key cybernetic notions that provide an information-flow framework within which to attack one aspect of the question of how a person thinks; i.e., the question of the information mechanisms and processes which underlie and are correlated with thinking.

After an introductory survey of the scope and ramifications of the information sciences, the cybernetic way of looking at information processing in the nervous system is examined so as to see in what sense it provides new and sharp tools of analysis for the neurophysiologist. With this as background, the problem of artificial intelligence is considered and with that the logical and linguistic difficulties in talking about the relationship between thinking and brain activity. An information-flow model of an artificial brain mechanism is described whose activity, it is argued,

is the correlate to activity such as perceiving, learning, thinking, knowing, etc.

This leads finally to a consideration of the impact of these notions on theoretical neurophysiology and its attempt to frame suitable hypotheses, and on epistemology which is concerned with the logical analysis of measures, methods, and techniques which can justify the activity of knowing.

14. U.S. - JAPAN SEMINAR ON MECHANICAL TRANSLATION: SUMMARY OF U.S. CONTRIBUTIONS
D. G. Hays
P-2892, March 1964, 11 pp.
AD 433 850
The summary examines MT as a branch of computational linguistics, reviews diverse theories in linguistics, notes current problems of descriptive linguistics, sketches computational problems, and notes administrative problems of international cooperation.

15. CYBERNETICS AND ITS DEVELOPMENT IN THE SOVIET UNION
R. Levien and M. E. Maron
RM-4156-PR, July 1964, 41 pp.
AD 602 705
Following a brief introduction, the Memorandum traces the birth of cybernetics and sketches its early development and its growth and emergence in the West. The second half of the Memorandum examines Soviet cybernetics.

16. REPORT OF A SUMMER SEMINAR ON COMPUTATIONAL LINGUISTICS
D. G. Hays
RM-3889-NSF, February 1964, 65 pp.
AD 431 868
(National Science Foundation)
The Seminar was held from July 8 through August 30, 1963, in Santa Monica. The staff taught elementary programming, linguistic theory, mathematical linguistics, and existing programs for linguistic operations to twenty professors of linguistics. The report includes a detailed schedule, an analysis and evaluation of the Seminar by the participants, and a list of all persons involved. The evaluation indicates that most participants, while attempting to obtain further education in this area, will meantime begin using computers in their own research and that many of them will also teach this subject - some in special courses, some in general linguistics courses, and others only to research students.

17. MACHINE TRANSLATION
K. E. Harper
In: T. A. Sebeok, ed., Soviet and European Linguistics, Current Trends in Linguistics, vol. I, Mouton & Co., The Hague, 1963, pp. 133-142.
After writing about 20 algorithms for different pairs of languages, Russian linguists turned their attention more to basic studies of language and development of theory. Harper scans the literature and summarizes a variety of recent work.

18. TECHNICAL WRITERS: EDUCATED OR TRAINED?
V. Peterson
P-3069, March 1965, 8 pp.
AD 612 380
Two possible innovations in the computer field which may influence the future of editors and technical writers are discussed; a typewriter-like console and a tablet. These changes should not affect either the responsibilities of the editor or his training and qualifications.

19. COMPUTATIONAL LINGUISTICS: BIBLIOGRAPHY, 1964
D. G. Hays and R. Ma
RM-4523-PR, March 1965, 53 pp.
AD 613 311
The bibliography cites 809 U.S. and foreign articles, reports, and books particularly relevant to the fields of computational linguistics and documentation, with selective coverage in the fields of classification theory, computation and programming, computers and hardware, nonnumerical application of computers, and psycholinguistics. In the area of linguistics, a fairly broad view of structural theory and semantics is taken without being exhaustive.

20. COMPUTATIONAL LINGUISTICS: BIBLIOGRAPHY, 1965
D. G. Hays, B. Henisz-Dostert, M. L. Rapp
RM-4986-PR, April 1966, 74 pp.
The bibliography cites 700 U.S. and foreign articles, reports, and books relevant to the fields of computational linguistics and documentation which appeared during late 1964 and throughout 1965. For a more detailed account of the coverage see 1-19.

21. ALCHEMY AND ARTIFICIAL INTELLIGENCE
H. L. Dreyfus
P-3244, December 1965, 98 pp.
AD 625 719
An examination of the difficulties of simulating cognitive processes on computers in the areas of game playing, problem solving, language translation, and pattern recognition reveals that the attempt to analyze intelligent behavior in digital computer language systematically excludes three fundamental human forms of information processing: fringe consciousness, essence/accident discrimination, and ambiguity tolerance. There are four distinct types of intelligent activity: 1. Associationistic (Irrelevance of meaning and context), 2. Non-formal (Dependent on meaning and context, which are not explicit), 3. Simple formal (Meanings completely explicit and context-independent), and 4. Complex formal (In principle, same as 3.; in practice, internally context-dependent, independent of external

context). Only areas 1. and 3. do not presuppose those human forms of information processing and can be programmed.

Significant developments in artificial intelligence in the remaining two areas must await computers which would combine their capacity for fast and accurate calculation with the short-cut processing made possible by the fringes of consciousness and ambiguity tolerance.

22. COMPUTATIONAL LINGUISTICS: RESEARCH IN PROGRESS AT THE RAND CORPORATION
D. G. Hays
P-3436, August 1966, 9 pp.
AD 637 929
RAND's linguistic research program is reviewed from its establishment in 1957. The current and projected work of the linguistics group is discussed in such areas as linguistic theory (in particular dependency and transformational), descriptive studies of Russian and to a lesser degree English, and the development of basic computing tools, especially the text and catalog system.

23. COMPUTATIONAL LINGUISTICS: BIBLIOGRAPHY, 1966
D. G. Hays, B. Henisz-Dostert, and M. L. Rapp
RM-5345-PR, April 1967, 89 pp.
The bibliography cites 914 U.S. and foreign articles, reports, and books relevant to the fields of computational linguistics and documentatior which appeared during late 1965 and throughout 1966. For a more detailed account of the coverage see 1-19.

24. READINGS IN AUTOMATIC LANGUAGE PROCESSING
D. G. Hays, ed.
(Unsponsored)
American Elsevier Publishing Company, Inc., New York, 1966, $10.00
The introduction (Hays) mentions the computer as reducer of linguistic data and deriver of conclusions from theoretical premises. It also mentions applications in social science, library work, translation, etc. It lists some techniques, said to unify the field: file management, textual input and output, dictionary lookup, and parsing. Among the papers in the volume are 3-3, 4.1-5, 4.2-5, and a brief expository paper on parsing by the editor.

25. INTRODUCTION TO COMPUTATIONAL LINGUISTICS
D. G. Hays
(Unsponsored)
American Elsevier Publishing Company, Inc., New York, 1967, $9.75
A textbook with chapters introducing computers and algorithms; storage structures; external storage; acquisition, storage, and presentation of textual data; dictionary lookup; parsing strategies; techniques for storing and using grammars; context-sensitive and transformational parsing; stratal conversion; concordances; techniques for linguistic research; documentation; and automatic translation.

2. Linguistic Theory

2.1. Dependency Theory

1. GROUPING AND DEPENDENCY THEORIES
D. G. Hays
RM-2646, September 1960, 20 pp.
(Previously published as P-1910, February 1960)
AD 250 237
In: H. P. Edmundson, ed., Proceedings of the National Symposium on Machine Translation, Prentice-Hall, Englewood Cliffs, N. J., 1961, pp. 258-266.
Immediate-constituent theory (phrase-structure theory) is based on a topology of grouping, whereas dependency theory uses a topology of trees, each minimal syntactic unit occupying a node in a tree. A concept of correspondence between the two kinds of structures is defined, and the two topologies are compared.

2. BASIC PRINCIPLES AND TECHNICAL VARIATIONS IN SENTENCE-STRUCTURE DETERMINATION
D. G. Hays
P-1984, May 1960, 22 pp.
In: Colin Cherry, ed., Information Theory, Butterworths, Washington, 1961, pp. 367-376.
Basic principles of the RAND method of sentence-structure determination include the isolation of grammatical detail from the structure of the computer program and postulation of a certain word-order rule, the rule of projectivity, that is realized in the program. Technical variations control the order of establishment of connections, the format of the grammar used in testing agreement, and other matters.

3. DEPENDENCY SYSTEMS AND PHRASE STRUCTURE SYSTEMS
H. Gaifman
P-2315, May 1961, 68 pp.
Information and Control, vol. 8, no. 3 (June 1965), pp. 304-337.
The formalism of dependency theory was first published here, together with proofs (i) that the RAND SSD routine is adequate for recognition of sentences produced by the formalism; (ii) that dependency and phrase-structure are weakly equivalent, since every language that has a finite grammar of one kind has a finite grammar of the other kind as well; and (iii) that the two theories are almost strongly equivalent, in the sense that all but a specified subclass of phrase-structure grammars correspond to dependency grammars, characterizing the same language and assigning corresponding structures to their sentences.

4. DEPENDENCY THEORY: A FORMALISM AND SOME OBSERVATIONS
D. G. Hays
RM-4087-PR, July 1964, 46 pp.
AD 602 648
Language, vol. 40, no. 4 (October-December 1964), pp. 511-524.

Dependency grammars characterize the class of context-free languages, assigning to each sentence of a characterized language a tree structure with minimal syntactic units at the nodes. Both production and recognition procedures are given. Either transformational or stratified linguistic systems can be constructed on the basis of dependency theory; more attention is given to the latter possibility. Semantic and psychological considerations are cited as motivating specific features of the theory, but they are no more necessary as justifications of this theory than for others.

5. AN ANNOTATED BIBLIOGRAPHY OF PUBLICATIONS ON DEPENDENCY THEORY
D. G. Hays
RM-4479-PR, March 1965, 14 pp.
AD 613 469

Seventy-one annotated citations in the field cover early works, formal analyses, parsing procedures, synthesis procedures, natural languages (English, Czech, French, and Russian) and discourse analysis.

2.2. Other Theory

1. AN EXPERIMENTAL STUDY OF AMBIGUITY AND CONTEXT
A. Kaplan
P-187, November 1950, 18 pp.
Mechanical Translation, vol. 2, no. 2 (November 1955), pp. 39-46.

Using English text, native speakers of English, and the criterion of distinguishing among dictionary definitions, Kaplan finds that relatively little context before or after a word suffices to eliminate most ambiguity.

2. THE QUANTIFICATION OF FUNCTIONAL LOAD: A LINGUISTIC PROBLEM
C. F. Hockett
RM-5168-PR, October 1966, 33 pp.
(Revised from P-2338, June 1961)

The function of a phonemic system is to keep the utterances of a language apart; distinctions that do more work are said to carry a heavier functional load. Hockett proposes criteria for indices of functional load which are satisfied by three measures of entropy (all used by Shannon in information theory); the three measures are in binits, in Shannons, and as relative entropy. The units analyzed can be phonemes, allophones, or components. The load carried by a contrast is non-negative; it is zero if the contrasted units are identical or in complementary distribution. Hockett's measure is the change in the entropy of the system if the contrasted phonemes are coalesced. He discusses some problems peculiar to the allophonic case.

3. SUPRASYNTACTICS
D. S. Worth
RM-3161-PR, June 1962, 21 pp.
AD 276 152

Suprasyntactic structures are those of negation, interrogation, and emphasis. The last of these is most fully treated in this report. Noting that stress on the governor in a combination (i.e., the head in a construction) is ambiguous, since it can mean emphasis either of the item itself or of the whole construction, Worth goes on to give numerous other examples of devices for emphasis in various languages and grammatical situations, fitting them all into a transformational model.

4. COMPUTERS AND COMPREHENSION
M. Kochen, D. M. MacKay, M. E. Maron, M. Scriven, and L. Uhr
RM-4065-PR, April 1964, 72 pp.
AD 437 589

Understanding entails construction of an internal model. The input to be understood must be handled not merely as symptomatic of a state of affairs, but as descriptive of a state. Comprehension is distinguished from learning. The highest level of comprehension involves ability to learn, to model the self and the interlocutor as well as the domain of discourse, and thus to engage in a fully reciprocal dialogue.

5. LINGUISTIC AND NON-LINGUISTIC "UNDERSTANDING" OF LINGUISTIC TOKENS
D. M. MacKay
RM-3892-PR, March 1964, 20 pp.
AD 432 308

This paper explores the distinction between recognizing an utterance, on the one hand, as a "symptom" of the state of affairs in or confronting its originator and, on the other hand, as a linguistic tool. It is suggested that although only the first kind of understanding is required to enable a computer to accept data and answer questions in verbal form, such ability is no guarantee of its comprehension of all aspects of language. To attain full linguistic comprehension, the program must also embody at least a "skeleton representation" of the linguistic context in which an utterance originates and from which it derives its linguistic significance as a goal-directed operator. By use of this richer model of its situation, a computer should be enabled to meet more sensitive tests of linguistic comprehension.

6. A SIMPLE PROOF OF A THEOREM ON SELF-SYNCHRONIZING AUTOMATA
D. M. Landi
P-3242-1, July 1966, 71 pp.
AD 636 097

A more direct, short proof is given for verifying Winograd's theorem that a finite state, completely specified automaton will resynchronize itself with probability 1. if and only if there exists a finite sequence of input letters called a universal synchronizer of the automaton.

7. SOME REMARKS ON ACCEPTABLE SETS OF NUMBERS
M. P. Schützenberger
P-3413, August 1966, 10 pp.
AD 637 123

Consideration is given to the application of negative results on the so-called acceptable sets of numbers to the case of push-down automata with the help of conventional analytic techniques. Two examples are given, and some related unsolved problems are discussed.

8. ON PRODUCTS OF FINITE DIMENSIONAL STOCHASTIC MATRICES
M. P. Schützenberger
P-3439, August 1966, 10 pp.
AD 638 847

This study presents and verifies a partial extension of a theorem of Wolfowitz on products of indecomposable, aperiodic stochastic matrices.

9. THE QUEENS GRAMMAR
S. L. Lieman
RM-5209-PR, January 1967, 48 pp.
AD 645 451

The famous eight-queens problem is presented and then generalized into a language over strings of integers. Various formalisms are shown to be able to describe this language, including arithmetical relations defined on each pair of integers in the string, programs for actually generating and parsing strings in the language, and a "relational grammar". It is shown that context-sensitive grammar and transformational grammar are likely to be capable of defining the language. The simplicity and elegance of the arithmetical formulation and the computer programs that are easily derived from it are contrasted to the counterintuitive properties of relational grammars, context-sensitive grammars and transformational grammars that define the same language. A natural language counterpart to the Queens Language is given, and new measures of linguistic adequacy are defined. The implications of these new measures of adequacy to efforts for finding appropriate linguistic formalisms is discussed.

3. Research Methods

1. RESEARCH PROCEDURES IN MACHINE TRANSLATION
D. G. Hays
RM-2916-PR, December 1961, 62 pp.
AD 268 941
In: Paul L. Garvin, ed., Natural Language and the Computer, McGraw-Hill Book Co., New York, 1963, pp. 183-214.

Processing natural language on a computer calls for precise, accurate, voluminous knowledge of the linguistic behavior of the speakers or authors whose utterances or writings are to be processed. That knowledge can be acquired by analysis of naturally produced or elicited texts, or from comments added to texts by editors. Syntax is described in terms of the dependency model, and indicators of syntactic connections (inflection, function words, occurrence order, and punctuation or intonation) are discussed. Under the heading of semantics, the problem of the unit of analysis is raised and the evidence that ambiguity exists is quoted. Calculation of the meaning of a sentence from the semantic properties of its parts and of their syntactic interrelations is discussed; the translation unit, or semantic unit, is the smallest linguistic unit that permits such calculation by simple rules. A scheme for automatic sentence-structure determination, the Cocke routine with an adaptation (suggested by Lecerf) to produce dependency structures, separates occurrence-order operations from all others, and obtains all possible structures for each sentence. Several plans for semantic recognition are mentioned, but none has been tested thoroughly, and all of them together still seem inadequate.

The text-based research procedures described are cyclic: some information about the language is available, and text is processed in accordance with that information, the results analyzed, and the new information used in processing more text. Standardization of equivalents, determination of contextual criteria for selection of equivalents, and improvement of syntactic analysis are considered as typical problems in linguistic research.

2. AN INTRODUCTION TO COMPUTATIONAL PROCEDURES IN LINGUISTIC RESEARCH
D. G. Hays
(Unsponsored)
In: Aldo Ghizzetti, ed., Automatic Translation of Languages, Papers presented at the NATO Summer School, Venice, 1962, Pergamon Press, Oxford, 1966, pp. 139-165

The mathematician begins with explicitly formulated rules or conventions and works out their consequences, which are the sentences of his formal languages. The native speaker of a natural language acquires a set of conventions by overt and covert learning; the linguist's task is to discover the covert conventions. Three methods of study are the distributional, the form-meaning, and the psychological; it is widely believed that the three methods should lead to the same results, but there are not yet proofs for this conjecture.

A method with posteditors is of the form-meaning type. Technical problems (worksheet formats, etc.) are discussed, especially for the case of multiple structures per sentence.

Analysis can use computers merely for making concordances, or for deeper parts of the work. Cross-tabulations, e.g., of inflectional suffix of dependent by inflectional suffix of governor, are suggested, with techniques for analysis of these frequency matrices.

Methods of fully automatic linguistic analysis have been proposed. Some suggestions of Harris, Garvin, and Lamb are reviewed, and a new procedure for determining dependencies is offered.

In conclusion, the notion of extrapolation from text to language is examined.

3. STUDIES IN MACHINE TRANSLATION--2: RESEARCH METHODOLOGY
H. P. Edmundson and D. G. Hays
RM-2060, December 1957, 24 pp (out of print)
AD 150 666
(Also published as P-1251, same date)
Mechanical Translation, vol. 5, no. 1 (July 1958), pp. 8-15.
Also in: David G. Hays, ed., Readings in Automatic Language Processing, American Elsevier Publishing Co., Inc., New York, 1966, pp. 137-148.

The procedure is cyclic. In each cycle, new text is compared with an existing glossary, and items in the text but not in the glossary are added to it; sentence-structure determination is attempted with an existing grammar, posteditors correct errors, and the errors are analyzed for the sake of improving the grammar; translation is attempted, postedited, and analyzed for the sake of improving the translation system.

4. ON THE VALUE OF A DEPENDENCY CONNECTION
D. G. Hays
RM-2712-AFOSR, January 1961, 26 pp.
AD 257 286
(Air Force Office of Scientific Research; AFOSR-TN-150)
In: 1961 International Conference on Machine Translation of Languages and Applied Language Analysis, Her Majesty's Stationery Office, London, 1962, pp. 577-592.
Also (in French) in: E. Delavenay, ed., Traduction Automatique et Linguistique Appliquée, Presses Universitaires de France, Paris, 1964, pp. 43-59.

In sentence-structure determination, values are numbers assigned to types of syntactic relations in such a way that connections of higher value are established in preference to connections of lower value. Given a text in which sentence structures are known (as by postediting), the values of some syntactic relations can be estimated by the following plan: assign value 1 to a relation provided no relation is known to have lower value; assign value 2 to a relation provided all relations known to have lower value are also known to have value 1; etc. The same procedure can be used for assigning adjectives to order classes, and for other similar purposes.

5. STUDIES IN MACHINE TRANSLATION--8: MANUAL FOR POSTEDITING RUSSIAN TEXT
H. P. Edmundson, K. E. Harper, D. G. Hays, and B. J. Scott
RM-2068, July 1960, 45 pp.
(Also published as P-1624, November 1959)
AD 249 679
Mechanical Translation, vol. 6 (November 1961), pp. 63-71.

Instructions for (i) choice of English equivalents, (ii) marking of English inflections, insertions, etc., and (iii) indication of dependency structure, all relative to a specific worksheet format and coding scheme. Includes some semantic and morphological suggestions about dependency analysis.

6. SOME MONTE CARLO ESTIMATES OF THE YULE DISTRIBUTION
H. A. Simon and T. A. Van Wormer
P-2599, July 1962, 19 pp.
AD 604 799
Behavioral Science, vol. 3, no. 3 (July 1963), pp. 203-210.

Some Monte Carlo simulations of the Yule process, viewed as a stochastic process for generating the frequency distributions of words in text. Part I presents data for the case that cannot be solved analytically in closed form, where the rate at which new words enter the text is variable. Part II compares Monte Carlo simulations with two sets of data in which both the empirical frequency distributions and the actual rate at which words enter the text are known.

7. THE LOGIC OF COGNATE RECOGNITION IN HISTORICAL LINGUISTICS
M. Kay
RM-4224-PR, September 1964, 25 pp.
AD 605 823

This paper presents a formalization of one step in the comparative method, that in which modern derivatives of prehistoric phonemes are recognized. Nhe basic assumption is that the words of a hypothetical prehistoric language should be constructed in such a way as to minimize the total number of phonemes in the language and of the statements that need to be made to account for the forms of the modern words. The theory is sufficiently specific to provide an algorithm for a computer program and provides a basis on which more efficient heuristic procedures might be built.

8. LANGUAGE IDENTIFICATION IN THE LIMIT
E. M. Gold
RM-4136-PR, July 1964, 40 pp.
AD 602 071

Information concerning a language is presented by a text, an unending succession of examples of strings in the language, or by informant, a source that tells whether or not successive strings are in the language. Decision procedures

generate successive guesses as to the identity of the language, which is said to be identified in the limit if these guesses are the same and correct after some finite time. Six variations of each of the two basic identification situations are considered. It is found that identification from informant is powerful enough to identify in the limit primitive recursive languages, which include context-free languages. Use of textual information, however, is so weak that, except for one special case, which appears to have no practical implications, not even regular languages are limiting identifiable.

9. THE TERMINATION OF CERTAIN ITERATIVE PROCESSES
R. M. Needham
RM-5188-PR, November 1966, 10 pp.
AD 641 667

This study shows that a certain class of iterative processes must terminate. The processes in question arise mostly in automatic classification, and are used for finding clusters of points or objects in a wide variety of spaces.

10. AUTOMATIC CLASSIFICATION IN LINGUISTICS
R. M. Needham
P-3500, December 1966, 15 pp.
AD 644 961

Prepared for publication in The Statistician, this Paper emphasizes the difference between methods appropriate for the classification of words and the usual forms of numerical taxonomy. Because words often have more than one meaning, linguistic classifications are overlapping. Techniques that yield a classificatory hierarchy are rankly inappropriate. The basic problem is not whether the right data have been chosen but how to handle the very large samples necessary. Randomization and construction of similarity matrices are likely to be unproductive. Independent searches for separate classes seem to work best. A nonmatrix technique employing a cohesion function and search for class profiles is outlined, which usually requires less than ten iterations for termination.

4. Computation

4.1. Codes and Formats

1. STUDIES IN MACHINE TRANSLATION--5: MANUAL FOR KEYPUNCHING RUSSIAN SCIENTIFIC TEXT
H. P. Edmundson, D. G. Hays, E. K. Renner, and R. I. Sutton
RM-2061, December 1957, 10 pp.
AD 150 668

Card layout for punching one form occurrence per card; keyboard layout for Cyrillic; punctuation code; and code for non-Cyrillic.

2. STUDIES IN MACHINE TRANSLATION--3: RESUME OF MACHINE CODES AND CARD FORMATS
H. P. Edmundson, D. G. Hays, and R. I. Sutton
RM-2064, August 1958, 33 pp.
(Also published as P-1352, same date)
AD 156 048

Codes for the Cyrillic alphabet, for text location, for punctuation, for non-Cyrillic matter in text, for glossary location, for English equivalents, etc. Card formats for text and glossary, assuming the use of punched-card equipment for many steps in data processing.

3. STUDIES IN MACHINE TRANSLATION--4: MANUAL FOR PRE-EDITING RUSSIAN SCIENTIFIC TEXT
H. P. Edmundson, D. G. Hays, E. K. Renner, and D. V. Mohr
RM-2065-1, December 1959 (revised June 1961), 38 pp.
AD 259 231

Instructions for sequence numbering and classification of non-Cyrillic material. An elaborated code for non-Cyrillic material is included, and full-line-per-card punching is assumed.

4. STUDIES IN MACHINE TRANSLATION--6: MANUAL FOR CODING RUSSIAN GRAMMAR
K. E. Harper, D. G. Hays, and D. V. Mohr
RM-2066-1, March 1958 (revised December 1960), 87 pp.
AD 253 635

Part-of-speech, case, number, gender, tense, case requirements of cardinal numbers, nouns, adjectives, adverbs, verbs, and participles, functions of short-form adjectives, aspect, degree, reflexivity, mood, person, identification of particles, prepositions, and conjunctions, classification of non-Cyrillic forms, and special classification with respect to prepositional relationships.

5. NATURAL LANGUAGE IN COMPUTER FORM
M. Kay and T. W. Ziehe
RM-4390-PR, February 1965, 81 pp.
AD 610 527
Also in: David G. Hays, ed., Readings in Automatic Language Processing, American Elsevier Publishing Co., Inc., New York, 1966, pp. 33-50.

This Memorandum describes a scheme for recording text in computer-usable form in such a way that all meaningful typographical distinctions are represented in a standard way. Provision is made for texts in different languages and different alphabets and for subsidiary material such as parallel translations and comments of interest to users and librarians. The basic set of encoding conventions is indefinitely extensible to accommodate new kinds of material.

Very large bodies of data require special facilities, and these have been provided

by embedding the text-encoding scheme in a general file-maintenance system. This provides for a comprehensive set of labels for different-sized units of text and makes it easy to retrieve any given unit. It also provides means of correcting and revising material in the file.

It is expected that files of computer-usable text will be built up in a variety of ways and, in particular, that the keypunching of text for this express purpose will become steadily less important. Computer programs are described which simplify conversion of text from these various sources into the standard format.

The final section discusses the problem of printing text which has been recorded in the standard format and describes a flexible program for doing this.

6. ACQUISITION, ARCHIVING AND INTERCHANGE
D. G. Hays
P-3497, December 1966, 24 pp.
AD 644 559

"Acquisition" is used to mean the conversion of information from a form designed for human consumption to one designed for machine input. This can be done using a standard keyboard device such as a typewriter (perhaps slightly modified, perhaps with a variable character set). Many different kinds of printed material can be encoded in ways that retain the important typographical distinctions, if the encoding conventions used for each item are stored with the item to enable it to be printed out for reading later. The RAND Text and Catalog System for the machinable archiving of printed text is described, and a format of still greater generality is outlined. A common encoding and formatting scheme is necessary to permit interchange of materials between archives, and it must be such as to allow the publication of high-quality extracts.
(See also RM-4390-PR, RM-4540-PR)

4.2. Programs & Systems

1. THE NATURE OF DATA IN LANGUAGE ANALYSIS
T. W. Ziehe and S. L. Marks
P-2197, January 1961, 18 pp.

Data used for and resulting from the analysis of natural language exhibit many different relationships; in the files that must be maintained for linguistic research, both the data and the relationships are subject to frequent modifications. Ziehe and Marks describe a generalized notation for the relationships, making it possible to treat the relationships themselves as data on which computer routines can operate without adaptation as the relationships change.

2. MIMIC: A TRANSLATOR FOR ENGLISH CODING
H. S. Kelly
P-1926, March 1960
In: H. P. Edmundson, ed., Proceedings of the National Symposium on Machine Translation, Prentice-Hall, Englewood Cliffs, N. J., 1961, pp. 325-334.

The translator and interpreter described here accept instructions, of limited variety but English-like form, simplifying the task of the linguist who must write many short programs for analysis, translation tests, and so on, altering each after a small amount of data has been processed. The system is designed to work on structured data, e.g., input text with grammar-code symbols, sentence-structure descriptions, and tentative equivalents attached.

3. GLOSSARY LOOKUP MADE EASY
H. S. Kelly and T. W. Ziehe
P-1909, February 1960, 16 pp.
In: H. P. Edmundson, ed., Proceedings of the National Symposium on Machine Translation, Prentice-Hall, Englewood Cliffs, N. J., 1961, pp. 325-334.

A glossary of forms occurring in a block of text is compiled by hash addressing. With this glossary in core storage, the glossary is read in two parts. In the first part, each item is compared with the contents of core storage, and a record of the items matched is compiled. In the second part, information about the matched items is selected by means of the record previously constructed and transferred to core, replacing the items originally there. Finally, this information is written out on tape in occurrence order in preparation for the next linguistic operation.

4. STUDIES IN MACHINE TRANSLATION--10: RUSSIAN SENTENCE-STRUCTURE DETERMINATION
D. G. Hays and T. W. Ziehe
RM-2538, April 1960, 78 pp.
AD 238 096

Compared with other programs available at the time, this one was relatively simple and easy to transfer from one problem (i.e., input language) to another. Dependency theory is sketched, and the isolation of word-order rules of precedence from all other grammatical rules is explained. Agreement tests are made using a table, with certain complexities to save space. Resultant grammatic types are the altered descriptions assigned to units when dependency connections are made. The special cases of conjunctions (which the program could handle), of ellipsis (for which a subroutine was to be written), and of relative clauses (which could be handled) are outlined in part. Punctuation and idiom recognition are mentioned. The routine consists of 2400 instruction words, running at about 650 text occurrences per minute on the IBM 704. Problems of structure revision--i.e., of backtracking when a partially completed structure is found to contain an error--are considered.

5. CONNECTABILITY CALCULATIONS, SYNTACTIC FUNCTIONS, AND RUSSIAN SYNTAX
D. G. Hays
(Paper completed at EURATOM)
Mechanical Translation, vol. 8, no. 1, (August 1964), pp. 32-51.
Also in: David G. Hays, ed., Readings in Automatic Language Processing, American Elsevier Publishing Co., Inc., New York, 1966, pp. 107-126.

Code matching is an alternative to table-lookup in tests of grammatical agreement. This plan requires elaborate descriptions of individual items (e.g., the words in a dictionary) but it avoids the use of large tables or complex routines for the tests. Development of the technique also leads to some clarification of the linguistic concepts of functions, exocentrism, and homography. A format for the description of Russian forms and a program for testing connectability by matching are described. Nine functions are recognized: subjective; first, second, and third complementary; first, second, and third auxiliary; modifying; and predicative. These are the dominative functions; another program must still be written for the coordinative functions: coordination, apposition, etc.

6. TRANSLATION OF ARTIFICIAL LANGUAGES BY COMPILER PROGRAMS
R. F. Rosin
P-1771, September 1959, 15 pp.

A program was written in GAT, the language of a compiler used at the University of Michigan, to translate FORTRAN-SIT programs into GAT programs. Rosin suggests several possible specifications that might be used for future translations of both artificial and natural languages.

7. EXPERIMENTS WITH A HEURISTIC COMPILER
H. A. Simon
P-2349, June 1961, 87 pp.
Journal of the Association for Computing Machinery, vol. 10, no. 4 (October 1963), pp. 493-506.

The experiments were performed during the construction of a compiler using heuristic problem-solving techniques such as those incorporated in the General Problem Solver. The experiments attempt to discover the problems of constructing more powerful programming languages and compilers and to test whether the task of writing a computer program can be considered a "problem" in the sense defined by the General Problem Solver.

8. A PARSING PROGRAM FOR CATEGORIAL GRAMMARS
M. Kay
RM-4283-PR, September 1964, 32 pp.
AD 605 822

A parsing procedure which finds all the structures of a sentence allowed by a categorial grammar has been written for the IBM 7090 computer. High speed and economy of storage are achieved by first putting grammatical tags into parenthesis-free form and then compiling them into lists of computer instructions.

9. PALL: RAND'S AUTOMATED ADDRESS BOOK
G. E. Bryan
RM-4258, September 1964, 91 pp.
AD 605 456

The PALL program ("Prints Address Lists and Labels") has been in use since early 1963 to produce the various pieces of paper - distribution lists, receipts, mailing labels, and order forms - required in the distribution of RAND publications. It maintains a master file of some 850 organizations approved to receive various categories of RAND reports. In design and operation, the program resembles a two-pass assembler and makes extensive use of modern string-handling, pointer, and dictionary techniques. Section I is a brief discussion of the origin, operation, and advantages of PALL. Section II contains instructions for users of the program, Section III, instructions for operators, Section IV describes program operations, and Section V discusses program formats.

10. ENDOCENTRIC CONSTRUCTIONS AND THE COCKE PARSING LOGIC
J. J. Robinson
P-3101, March 1965, 20 pp.
AD 614 896

Cocke's Parsing Logic is described with a flowchart. When it is applied to a string like "All the old men on the corner," 16 partial parsings result. Robinson notes that the rules and parsing logic can be made to attach all modifiers on one side before beginning with those on the other side of the head. She gives some other methods tending to avoid waste of computing effort during parsing.

11. LARGE FILES IN LINGUISTIC COMPUTING
M. Kay
P-3136, May 1965, 18 pp.
AD 615 301

The text and catalog system more fully introduced in RM-4390-PR (the entry 4.1-5 above) is here described briefly and exemplified with applications to a Japanese-English dictionary and Japanese text. In particular, a text encoding scheme for the set of idiograms adapted from Chinese and used in Japanese is presented.

12. THE CATALOG: A FLEXIBLE DATA STRUCTURE FOR MAGNETIC TAPE
M. Kay and T. W. Ziehe
RM-4645-PR, October 1965, 25 pp.
AD 623 938
AFIPS Conference Proceedings, Vol. 27, Part 1 (1965), pp. 283-291.

This Memorandum outlines a generalized storage scheme for large files of highly structured data, or catalogs, and describes their realization on magnetic tape. Each datum, which may be large or small, is assigned to one of a number of data classes of which a user may define any number. The overall organization of a catalog is given by a map which imposes a tree structure on the set of data classes. Catalogs may participate as individual data in other catalogs so that the structure of a file may be recursive. A flexible addressing scheme not only facilitates the retrieval of data and sets of data in response to

easily formulated requests, but provides a basis for updating catalogs. General procedures called transformations can be used to derive catalogs with new structures from existing catalogs.

13. THE CATALOG INPUT/OUTPUT SYSTEM
M. Kay, F. Valadez, and T. W. Ziehe
RM-4540-PR, March 1966, 64 pp.

This Memorandum completely defines the format used for catalogs on magnetic tape and describes in detail the routines of the Catalog Input/Output System. Catalog maps, catalog data, and tape labels are written as logical records in a specially designed blocking format. Beginnings and ends of blocks, physical tapes, and catalogs are explicitly marked. The Catalog Input/Output System offers a variety of input/output unit-control operations in addition to the commands for reading and writing data in catalog format. The reading and writing commands are implemented on three levels: Level III handles individual data; Level II reads and writes logical records; and Level I processes blocks of information. Anyone planning to use the system should note that the structure processed on Levels I and II is ordinarily not of interest to the user. Only the simpler Level III structure reflects the user's view of his file in most applications. The Catalog Input/Output System is not intended to supplant specialized programs for extracting the contents of data and analyzing them in accordance with the user's special problems. For this reason, the user will ordinarily attempt to put one unanalyzable item of information into each datum, and prepare special programs, as few as possible, for content processing.

14. THE TABULAR PARSER: A PARSING PROGRAM FOR PHRASE STRUCTURE
M. Kay
RM-4933-PR, July 1966, 51 pp.
AD 635 934

A parsing program for the IBM 7040/44 computers is described which finds all phrase and dependency structures assigned to a sentence by a given grammar. The grammar is prepared for the machine in the form of a set of rule tables which constitute a rich, terse, and perspicuous grammatical notation.

15. COMPUTER ROUTINES TO READ NATURAL TEXT WITH COMPLEX FORMATS
P. Graves, D. G. Hays, M. Kay, and T. W. Ziehe
RM-4920-PR, August 1966, 133 pp.
AD 637 303

This Memorandum describes a set of subroutines for the IBM 7040/44 computers for reading textual material with complex formats and coding conventions--questionnaires, library catalog cards, etc.--from any external medium into the high-speed store of the machine. Different kinds of information in the input are recognized by explicit markers, position on the line or page, or syntactic clues given by other items. Less complex material requires only a portion of the system. Information may be recoded according to the user's conventions before being delivered to his program. The routines may be called from either FORTRAN- or MAP-coded programs.

16. THE CATALOG: A FLEXIBLE STRUCTURE FOR DATA STORAGE
T. W. Ziehe
P-3476, November 1966, 22 pp.
AD 642 368

An over-all view of the RAND text and catalog system described in RM-4390-PR for processing natural language text for computer storage in large files is presented. The system provides freedom to reorganize the files at will. Organization is based on the dependency tree. Data within any one class may be encoded in any form, disregarding the rest of the file, and the nature of the coding is indicated on the catalog map (index). Catalogs can be nested, and each catalog may be structured separately. A valuable space-saving device is the null datum, in which headings without any data fulfill their organizing role without occupying any physical space. Computer programs convert text into its internal representation. Formats have been defined for 7-bit magnetic tape and 36-bit-word core storage, and others will be added. Catalog transformation schemes are also planned for the future. (Presented at the Conference on Computers in Humanistic Research, Texas A. & M., November 1966.)

17. COLLECT: A PROGRAM FOR THE RETRIEVAL OF GRAMMATICAL INFORMATION FROM ANNOTATED TEXT
M. Kay and T. D. Taft
RM-5243, January 1967, 34 pp.

COLLECT is a system for searching a File of Russian sentences annotated with syntactic (and, eventually, morphological) information. The File consists of sentences; the user writes a description of the class of sentences he wishes to examine, and COLLECT edits his request, searches the File (taking several requests simultaneously, if desired), and prints a report. A sentence description consists of descriptions of word-form occurrences -- their spellings, grammatical properties, etc. -- and their relationships to one another. Small units of descriptions can be composed by conjunction or disjunction, and can be negated.

Although the system was written for this specific File and the kinds of searches linguists can be expected to make, it is flexible enough to permit addition of new kinds of information to the File, enrichment of the request terminology allowed, and modification of report formats. It is also representative of a class of retrieval programs needed for many file-search tasks.

The COLLECT system has been written for the IBM 7044 computer.

4.3. Program Languages

1. THE EVOLUTION OF CONCEPTS AND LANGUAGES OF COMPUTING
R. D. Elbourn and W. H. Ware
P-2526, December 1961, 8 pp.
Proceedings of the IRE, vol. 50, no. 5 (May 1962), pp. 1059-1066.

A review of the evolution of programming languages from the time when all programming was done in machine languages, through symbolic coding systems, interpreters, assemblers, generators, and compilers, to the recently developed list-processing languages. These languages are then applied to game playing, problem solving, and theorem proving. Behavior and biological modeling are described. Finally, in anticipation of extending the capability of computers to accept, use, and generate natural languages, this paper concludes with an introduction to some of the contemporary work on formal language theory, including a discussion of six families of abstract languages and their practical implementation.

2. AN EXPERIMENTAL SYNTAX-DIRECTED DATA STRUCTURE LANGUAGE
R. K. Lindsay, T. W. Pratt, and K. M. Shavor
P-3112, April 1965, 48 pp.
AD 614 782

Programmers developing systems of the complexity required in artificial intelligence research are frequently hindered by the rigid programming languages available and the time-consuming task of implementing new languages. AMOS (for associative memory organizing system) provides a flexible means to structure data and experiment with the syntactic forms of program statements while lessening the implementation bottleneck. AMOS is a syntax-directed compiler used to define languages for constructing a variety of data organizations of which FORTRAN-like arrays and IPL-like list structures are special cases. This research explores the use of syntactic descriptions which are not Backus Normal Form grammars and provides means for defining two-dimensional languages as well as the usual linear type. In order to facilitate implementation, the system may be conveniently imbedded in any monitor system of common design; AMOS operations are manipulations within high-speed storage only.

5. The Russian Language

5.1. Text and Glossaries

1. STUDIES IN MACHINE TRANSLATION: BIBLIOGRAPHY OF RUSSIAN SCIENTIFIC ARTICLES
K. E. Harper, D. G. Hays, A. S. Kozak, and B. J. Scott
RM-3610-1, June 1963 (revised November 1966), 125 pp.

A catalogue of Russian text available on magnetic tape at RAND. Thirty-one corpora are included:

Subject field	Corpora	Pages	Occurrences
Physics	12	1507	373,329
Mathematics	4	471	108,124
Astrobotany	8	936	260,188
Cybernetics	6	645	208,565
Physiology	1	378	111,530
Total	31	3937	1,061,736

The physiology text was keypunched by a research project at Georgetown University several years ago. For each article, the bibliography gives title, author, and citation, together with length (in pages and occurrences) and RAND identification number.

2. A GLOSSARY OF RUSSIAN PHYSICS ON PUNCHED CARDS
K. E. Harper, D. G. Hays, and A. Koutsoudas
P-1241, December 1957
Mechanical Translation, vol. 4, no. 1/2 (November 1957), p. 1.

In 1957, a glossary of 6,000 Russian forms, prepared cooperatively by the University of Michigan and RAND, was offered to qualified researchers.

3. STUDIES IN MACHINE TRANSLATION--12: A GLOSSARY OF RUSSIAN PHYSICS
A. S. Kozak, C. H. Smith, and members of the RAND ALDP group
RM-2655, October 1960, 311 pp.
AD 252 609

A list of the 23,146 distinct forms of 6,847 words that occur in 265,417 occurrences of Russian physics text. With each form are shown the English equivalents, grammatic properties, and occurrence frequency derived from the text by the RAND cyclic procedure. The forms first observed in the last batch of 30,000 occurrences are included, but no postediting results from that corpus are embodied in the equivalents and grammar-code symbols.

4. HIGH FREQUENCY WORDS AND OCCURRENCE FORMS IN RUSSIAN PHYSICS
A. S. Kozak
RM-3383-PR, October 1962, 102 pp.
AD 288 133

This report lists the 1026 highest-frequency words and the 1013 highest-frequency forms in the RAND physics glossary. Each collection is listed by frequency and also alphabetically. (Non-Cyrillic items are excluded.) Comparison with Josselson's Russian word count, based on general prose, shows that 150 of his 204 most common words do not appear among the 204 most common words in the RAND text. Moreover, 36 of these words do not occur in the quarter-million occurrence sample of physics text. The 1026 words have frequency 35 or higher; their total frequency in the text is

202,115, or 86.30% of the occurrences. The 1013 forms also have frequency 35 or higher, but their total frequency is only 146,308, or 62.47% of all occurrences.

5. A FILE OF RUSSIAN TEXT WITH SYNTACTIC ANNOTATIONS
D. G. Hays, D. B. Gottshall and A. S. Kozak
RM-5194-RADC, January 1967, 28 pp. (Rome Air Development Center)

The File described here is a corpus of more than a million running words of Russian text, mostly coming from the journal literature of physics, mathematics, astrobotany, and physiology, but syntactic examples from a large grammar are included to provide variety for syntactic and possibly semantic studies. All sentences have been annotated with syntactic structures, using the notation of dependency theory. Each sentence is shown to have an independent word; other words serve specified functions for specified governors. The File is recorded on magnetic tape as a "catalog."

5.2. Syntax

1. ORDER OF SUBJECT AND OBJECT IN SCIENTIFIC RUSSIAN WHEN OTHER DIFFERENTIA ARE LACKING
D. G. Hays
P-1632, March 1959, 6 pp.
Mechanical Translation, vol. 5, no. 3 (December 1958), pp. 111-113.

In a sample of 22,000 occurrences of Russian physics text, 56 instances of true ambiguity were found; in these instances, no morphological properties of subject or object (head word or modifiers) reduced the ambiguity of either. In these instances, the subject was before the verb 52 times, the object always after the verb, and the four cases of subject-after-verb seemed idiomatic, involving frozen combinations of verb and object.

2. A RUSSIAN STRUCTURE FOR COMPARISON
D. G. Hays and B. J. Scott
P-1720, June 1959

Three fixed combinations, tot zhe...chto i, takoj zhe...chto i, and takoj zhe... kak i, all translated the same...as, are used to draw comparisons between objects, properties, or circumstances. Using postedited text (120,000 occurrences), Hays and Scott found 27 sentences containing the constructions, and posited three transformations to account for their use. Techniques for identification of the transformations and points of application are proposed.

3. GOVERNORS OF THE CONJUNCTION CHTO
J. H. Pustula
P-1978, February 1960, 16 pp.

The Russian subordinate conjunction chto (that) introduces noun clauses. In the RAND corpus of about 240,000 occurrences, 1,466 clauses with chto were governed by 85 different verbs, present participles, short-form past participles and adjectives, and nouns, and by 4 idioms. These governors are listed, each with frequency, and classified.

4. SOME LINGUISTIC PROBLEMS OF RUSSIAN GRAPHIC ABBREVIATIONS
W. A. Stewart
P-2206, January 1961, 14 pp.

Abbreviations are mostly undeclinable, and therefore devoid of the morphological indicators generally useful in sentence-structure determination. They are sometimes reductions of single words, sometimes of phrases. They exhibit the structural features of the natural language in which they are used, and present more or less the same problems, with the above-noted exceptions.

5. STUDIES IN RUSSIAN MORPHOLOGY--I: THE SUFFIX -AGA
D. S. Worth
RM-3235-PR, August 1962, 45 pp.
AD 281 850
Scando-Slavica, vol. 10 (1964), pp. 174-193.

About two dozen words in contemporary standard Russian are formed by adding the suffix -aga to nominal, adjectival, and verbal stems. The report considers the formation of derivational stems, the consonantal and vocalic alternations which occur when the suffix is added, and the gender of the derived word. Alterations in meaning are discussed, and Worth compares these items with another dozen ending in -aga but not with the suffix (i.e., the ending is homophonous with the suffix). He asserts that the two classes are brought together by speakers of the language, and treated by them as falling into two classes: (i) the referent is usually human and the suffix is expressive, sometime pejorative, sometimes favorable; (ii) the referent is usually nonhuman and the suffix is not expressive.

6. STUDIES ON PREDICATION IN RUSSIAN--I: PREDICATIVE CASE, SHORT FORM ADJECTIVES AND PREDICATIVES
H. Birnbaum
RM-3774-PR, January 1964, 86 pp.
AD 429 524
Abridged in: California Slavic Studies, vol. IV, University of California Press, Berkeley and Los Angeles (in press).

This is the first, theoretical part of a series of studies on certain aspects of predication (predicative function) and its formal expression in Contemporary Standard Russian (CSR). This series is primarily concerned with such instances where in Russian the predicate is denoted by means other than a finite verb, i.e., by infinite verbal forms (infinitives and gerunds) as well as nonverbal forms, including the zero morpheme of the copula verb. In its sequels the series makes use of automatic and semi-automatic procedures applied on corpora of analyzed Russian physics texts available at the RAND Linguistic Research Project, in addition to evidence from other Russian linguistic materials.

7. COMPLEMENTATION IN RUSSIAN: THEORY AND APPLICATION
A. S. Kozak
RM-4582-PR, September 1965, 42 pp.
AD 620 932

Kozak reviews the concept of complementation in Russian, from both the Soviet point of view and that adopted at RAND for compilation of a file of scientific Russian text with syntactic annotations. Models for different patterns of complementation are presented.

8. A DEEP INDEX OF DERIVATIONAL MORPHOLOGY
D. S. Worth
RM-5143-PR, September 1966, 37 pp.
AD 640 962

A projected "Deep Index" designed to gather, file and retrieve, with the assistance of the "catalog" system, a large quantity of information concerning the extant literature on the derivational morphology of Russian is described.

The Memorandum describes the type of literature being abstracted, the classification of information into discrete categories for retrieval, the steps taken to prepare data for conversion to machine format, and the transliteration and abbreviation conventions used in the Index. An example of Russian scholarly text is appended, together with the format in which this text appears after processing.

9. TOWARD EXPLOITATION OF A FILE OF RUSSIAN TEXT WITH SYNTACTIC ANNOTATIONS
D. G. Hays and D. S. Worth
RM-5252-RADC, January 1967, 45 pp.
(Rome Air Development Center)

After summary descriptions of a File of Russian text with syntactic annotations and the system of computer programs called COLLECT usable in searching the File, this Memorandum discusses methods of automatic (statistical) classification usable in reducing data obtained from the File, the use of parsing programs in linguistic research, and two of the many kinds of syntactic questions that might be answered by research based on the File: modal constructions and sentential apposition.

10. STUDIES IN RUSSIAN MORPHOLOGY: II. VOWEL-ZERO ALTERNATIONS IN DERIVATION
D. S. Worth
RM-5223-PR, May 1967, 40 pp.

In one type of vowel-zero alternation, a flexional-level vowel-zero morphophoneme is stabilized as either a full vowel or a zero in the course of derivation. In another type, a full vowel in the derivational base alternates with zero in the derived word, or vice-versa. Worth suggests the existence of a vowel-zero morphophoneme on the derivational level as an explanation for certain of these alternations. As background, he sketches a generative framework within which to consider specific techniques of word formation, and briefly surveys flexional vowel-zero alternations.

5.3. Semantics

1. PAIRS OF RUSSIAN WORDS WITH HIGH CORRELATION
D. G. Hays
P-1218, November 1957, 69 pp.

Using the material prepared by H. H. Josselson for The Russian Word Count, co-occurrence (on the same page) was tallied for every possible pair of words on a list of 200 common words in 4037 pages of text. The 2731 most highly correlated pairs are listed.

2. PROCEDURES FOR THE DETERMINATION OF DISTRIBUTIONAL CLASSES
K. E. Harper
RM-2713-AFOSR, January 1961, 24 pp.
AD 257 287
(Air Force Office of Scientific Research; AFOSR-TN-149)
In: 1961 International Conference on Machine Translation of Languages and Applied Language Analysis, Her Majesty's Stationery Office, London, 1962, pp. 687-700.

A distributional class is a collection of words each of which can bear a specified syntactic relation to any member of some other set. The second set used in the definition may be a morphological or syntactic class, or one formed a priori, that is, an intuitively relevant class--such as the class of "abstract" nouns--hence, perhaps, a semantic class. Some distributional classes are examined: that of verbs having only animate actor nouns as dependents, that of governors of nouns naming physical particles, and that of governors of rezkij (sharp). Several problems raised by the procedure are considered.

3. TRANSFORMATIONAL CRITERIA FOR THE CLASSIFICATION OF PREDICATIVE GENITIVE CONSTRUCTIONS IN RUSSIAN
D. S. Worth
RM-2714-AFOSR, January 1961, 20 pp.
AD 257 288
(Air Force Office of Scientific Research; AFOSR-TN-148)
In: 1961 International Conference on Machine Translation of Languages and Applied Language Analysis, Her Majesty's Stationery Office, London, 1962, pp. 725-737.

Some Russian constructions containing genitive substantives must also include a modifier, such as an adjective or a genitive substantive. For example, "Sapog bol'shogo razmera" ("A boot of large size") cannot be reduced to "Sapog razmera" ("A boot of size"). Several varieties of such constructions are determined by transformational possibilities; Worth finds that every variety contains or implies a predication: "A boot, of which the size is large."

4. MACHINE TRANSLATION OF RUSSIAN PREPOSITIONS
K. E. Harper
P-1941, May 1960, 27 pp.

Using postedited text, Harper classified the governors and dependents of Russian

prepositions according to their effect on translation of the preposition. Some idioms are recognized (e.g., v zaklyuchenie (in conclusion), and some fixed combinations (e.g., sostoyat' v (to consist of)) that cannot be called idioms in the narrow sense because word order is not fixed. 42 prepositions that occurred about 650 times in about 240,000 occurrences had only one equivalent each. 12 prepositions, occurring about 12,000 times in about 120,000 occurrences, could be translated accurately in the sample text upon identification of 60 fixed combinations and a dozen word classes.

5. THE POSITION OF PREPOSITIONAL PHRASES IN RUSSIAN
K. E. Harper
RM-3625-PR, May 1963, 17 pp.
AD 405 561
In: Mechanical Translation, vol. 8, no. 1 (August 1964), pp. 5-10.

In automatic sentence-structure determination, selection of the governors of prepositional phrases is reputed to be exceedingly difficult. Harper first sets aside combinations in which the preposition is strongly governed, i.e., those in which there is a special relation between the preposition and its governor; these combinations can be listed, and the preposition is almost always dependent on the corresponding governor when the two occur together in a clause--but only about 1 occurrence in 5 is strongly governed. Examining various subsamples of the RAND-parsed Russian physics text, he finds: (i) Verbs and nouns are the most frequent governors of prepositions, adjectives about 1%, other parts of speech practically not at all. (ii) When a noun and a verb compete for the preposition, the noun itself depending on the verb, then (a) projectivity makes the preposition depend on the nearest element in N-V-P and P-V-N, (b) by observation, the preposition almost always depends on the verb in P-N-V and V-P-N, and the exceptions are syntactically recognizable, and (c) in the order N-P-V or V-N-P, the noun governs the preposition in about 90% of the occurrences. (iii) In case c above, when the verb is governor, usually either the prepositional phrase is recognizably adverbial (e.g., in reality), or the noun that does not govern the preposition is in an adverbial construction (e.g., itself the object of an adverbial prepositional phrase). There remains for further study the case of a prepositional phrase following a sequence of nouns, i.e., N-N-...-N-P, where each noun depends on the one before.

6. PREPOSITIONAL PHRASES AND AUTOMATIC PARSING
K. E. Harper
RM-4383-PR, February 1965, 27 pp.
AD 611 306

A concordance was prepared for machine-processed Russian text, showing the syntactic governors of prepositional phrases when the word preceding the preposition was a noun (P-1=N). The concordance showed that under this condition, Russian prepositions strongly tend to serve either an adnominal function or an adverbial function. These properties are not inherent in the individual preposition, but are environmentally determined. The resultant grouping of prepositions is presented as information that can be used in a program for sentence-structure determination.

7. MEASUREMENT OF SIMILARITY BETWEEN NOUNS
K. E. Harper
RM-4532-PR, May 1965, 21 pp.
AD 615 319

A study was made of the degree of "similarity" between pairs of Russian nouns, as expressed by their tendency to occur in sentences with identical words in identical syntactic relationships. A similarity matrix was prepared for forty nouns; for each pair of nouns, the number of shared (i) adjective dependents, (ii) noun dependents, and (iii) noun governors was automatically retrieved from machine-processed text. The similarity coefficient for each pair was defined to be the ratio of the total of such shared words to the product of the frequencies of the two nouns in the text. The 780 pairs were ranked according to this coefficient. The text comprised 120,000 running words of physics text processed at The RAND Corporation; the frequencies of occurrence of the forty nouns in this text ranged from 42 to 328.

The results suggest that the sample of text is of sufficient size to be useful for the intended purpose. Many noun pairs with similar properties (synonymy, antonymy, derivation from distributionally similar verbs, etc.) are characterized by high similarity coefficients; the converse is not observed. The relevance of various syntactic relationships as criteria for measurement is discussed.

8. STUDIES IN INTER-SENTENCE CONNECTION
K. E. Harper
RM-4828-PR, December 1965, 16 pp.
AD 626 572

Portions of the Russian physics text processed at The RAND Corporation were subjected to systematic analysis to determine the extent of repetition in adjacent sentences. Recurrence of the words in all pairs of contiguous sentences in the text (2,467 pairs) was recorded in a machine printout; sentence pairs for which word recurrence was not automatically recorded were visually inspected for other types of recurrence (through lexical stems, pronouns, synonyms, and paraphrases). The extent of the different types of recurrence is reported, and features of the recurring items are discussed. Sentence pairs characterized by nonrecurrence (12 percent of the total) are examined, and the relevance of inter-sentence recurrence and nonrecurrence to automatic syntactic analysis and abstracting is suggested.

9. STUDIES ON PREDICATION IN RUSSIAN-II: ON THE PREDICATIVE USE OF THE RUSSIAN INFINITIVE
H. Birnbaum
RM-4477-PR, December 1965, 251 pp.
AD 626 207

Birnbaum here continues the study begun in RM-3774 (the entry 5.2-6 above). He examines the predicative use of the Russian infinitive against the background of the entire range of semantic functions it serves in contemporary standard Russian. In the colloquial language, the infinitive increasingly competes with the finite verb. Among the uses examined are: semi-predicative, wherein object infinitives occur in combination with the dative complement; predicative infinitives in one and two membered sentences and subordinate clause equivalent phrases; constructions with by. Methods of transformational analysis of these predicative constructions are briefly discussed and exemplified. The Memorandum concludes with a few suggestions about automatic translation.

10. SOME COMBINATORIAL PROPERTIES OF RUSSIAN NOUNS
K. E. Harper
RM-5077-PR, September 1966, 54 pp.
AD 638 924

The study, based on 280 nouns which occurred at least 50 times in 180,000 of parsed Russian physics text, investigates the "valence" of nouns (tendency to occur as the syntactic governor of a noun in the genitive case). The valence of most nouns is intermediate. Factors influencing valence are: 1. a noun's position in a noun string (noun strings of length one or two account for almost 90 per cent of all noun strings), 2. the potential of the dependent genitive noun to form an adjective (partial interchangeability is observed), 3. valence is reduced in certain stable lexical combinations, 4. valence is decreased in non-initial occurrences of a noun in an article. Among the nouns of high valence are deverbative nouns and names of physical properties (abstract nouns), and among those of low valence are abbreviations for units of measurement, proper names, nouns employed in set phrases, and concrete nouns. Considerable variation in valence was observed for nouns of the same semantic or derivational classes, but sample studies suggest that these differences are well correlated with certain other morphological or syntactic features. In general, the results suggest that (1) for some nouns, valence may be attributed to certain inherent semantic properties, and (2) for other nouns, valence may be highly correlated with other "contextual" features both syntactic and lexical.

6. The English Language

1. PRELIMINARY CODES AND RULES FOR THE AUTOMATIC PARSING OF ENGLISH
J. J. Robinson
RM-3339-PR, December 1962, 149 pp.
AD 295 651

The author relies on a computer program that obtains all possible immediate-constituent structures for a sentence, given a grammar. The subject of this paper is a preliminary codification of English grammar, including the establishment of categories and rules for combinations of categories, an encoding scheme, and an analysis of the results obtained when the parsing program and grammar were applied to selected, but not always simple, sentences. Over 500 condensed rules, equivalent to more than 2,500 rules written without condensation, are listed. 65 structures for 16 sentences, ranging up to 37 occurrences in length, are presented by reproduction of machine listings.

2. PARSE: A SYSTEM FOR AUTOMATIC SYNTACTIC ANALYSIS OF ENGLISH TEXT - PARTS I & II
J. J. Robinson and S. Marks
RM-4654-PR, September 1965, 457 pp.
AD 621 310 and AD 621 311

This Memorandum explains PARSE, the RAND system for the automatic syntactic analysis of English sentences. The system can analyze and label the structures of a variety of sentences, including those with coordinate, subordinate, relative, indicative, and interrogative clauses.

PARSE consists of three major components: 1) a glossary of English word-forms for which grammar codes have been designated; 2) a table of grammatical rules listing the permitted combinations of adjacent grammar codes and assigning a code to the combination; and 3) a program for applying the rules systematically and exhaustively to English text coded from the glossary. The computer program was submitted to SHARE during 1965, together with a table of rules and a glossary.

The Memorandum is bound in two parts; part I gives linguistic details and describes the programs; part II contains the glossary of word-forms.

The system employes a parsing logic, devised by John Cocke of IBM, which disengages the grammar from the routines that apply it to text, so that changes in grammar will not affect the program. This feature is not only highly desirable; it appears to be essential in view of the complex task of developing rules adequately detailed to label intricate patterns of actual texts.

The logic also applies the grammar exhaustively: every construction recognized by the grammar is displayed for every string submitted for analysis. This feature provides data for evaluating the grammar and locating the points where modifications should be made. It also reveals numerous unsuspected syntactic and semantic ambiguities in many of the sentences analyzed, and furnishes a framework for determining how sentences are understood and disambiguated through linguistic and extra-linguistic contexts.

3. SOME ASPECTS OF THE THEMATIC ORGANIZATION OF THE ENGLISH CLAUSE
M. A. K. Halliday
RM-5224-PR, January 1967, 28 pp.
AD 646 663

Halliday's discussion is in terms of his "systemic" grammar (roughly corresponding to "deep" analysis) at the level of the clause. From the point of view of the organization of a clause as a message (thematic organization), three types of choices are distinguished as principal: 1. information, realized phonologically by intonation features and concerned with the given-new, 'given' being what the speaker can derive from the preceding discourse, 2. thematization, realized by the sequence of elements in the clause, and dealing with theme and non-theme, or theme and rheme, theme being assigned first position in a sequence; and 3. identification, realized by certain structural features (mainly the indication of information focus by intonation) and dealing with the known-unknown in 'equative' or identifying clauses. These sets of choices are closely bound up with each other and with the other syntactic selections that characterize the clause in English.

7. Retrieval & Indexing

1. ON RELEVANCE, PROBABILISTIC INDEXING AND INFORMATION RETRIEVAL
M. E. Maron and J. L. Kuhns
(Paper written at Thompson Ramo Wooldridge, Inc.)
Journal of the Association for Computing Machinery, vol. 7, no. 3 (July 1960), pp. 216-244.

This paper reports on a novel technique for literature indexing and searching in a mechanized library system. The notion of relevance is taken as the key concept in the theory of information retrieval and a comparative concept of relevance is explicated in terms of the theory of probability. The resulting technique, called "Probabilistic Indexing," allows a computing machine, given a request for information, to make a statistical inference and derive a number (called the "relevance number") for each document, which is a measure of the probability that the document will satisfy the given request. The result of a search is an ordered list of those documents which satisfy the request ranked according to their probable relevance.

The paper goes on to show that whereas in a conventional library system the cross-referencing ("see" and "see also") is based solely on the "semantical closeness" between index terms, statistical measures of closeness between index terms can be defined and computed. Thus, given an arbitrary request consisting of one (or many) index term(s), a machine can elaborate on it to increase the probability of selecting relevant documents that would not otherwise have been selected. Finally, the paper suggests an interpretation of the whole library problem as one where the request is considered as a clue on the basis of which the library system makes a concatenated statistical inference in order to provide as an output an ordered list of those documents which most probably satisfy the information needs of the user.

2. AUTOMATIC INDEXING: AN EXPERIMENTAL INQUIRY
M. E. Maron
RM-2601, August 1960, 35 pp.
(Also published as P-2180, February 1961)
AD 245 175
Journal of the Association for Computing Machinery, vol. 8, no. 3 (July 1961), pp. 404-417.

An empirical test of a statistical technique is reported. The technique uses differential word frequencies in preestablished categories.

3. A LOGICIAN'S VIEW OF LANGUAGE DATA PROCESSING
M. E. Maron
P-2279, August 1961, 47 pp.
In: Paul L. Garvin, ed., Natural Language and the Computer, McGraw-Hill Book Co., New York, 1963, pp. 128-150.

The nature of logic is described in relation to both natural language and the problem of language-data processing. Marons analyzes the problem of information identification and retrieval from this point of view.

4. INFORMATION RETRIEVAL: A LOOK AT THE LOGICAL FRAMEWORK AND SOME NEW CONCEPTS
M. E. Maron
P-2455, October 1961, 37 pp.

Weighted index tags allow for computing a measure of the relevance of a document to an information request. Statistical techniques aid in automatically relating and associating documents on the basis of their subject content. However, further work is needed (i) to allow the machine to identify documents on the basis of its "knowledge" of the complete text and (ii) for better understanding of the processes behind intelligent problem solving.

5. PROBABILITY AND THE LIBRARY PROBLEM
M. E. Maron
P-2471, January 1962, 21 pp.
Behavioral Science, vol. 8, no. 3 (July 1963), pp. 250-257.

Automatic information retrieval systems are needed because of the growing rate of scientific publication. Besides ordinary libraries, such systems can be used in patent offices, medical practice, law, military intelligence, etc. The system should process incoming documents automatically, disseminate them on standing orders, and store the content of the documents for later requests. Some machines designed for the purpose are mentioned. The problem of determining the relevance of a document to a request is analyzed probabilistically.

6. A PROPOSAL FOR THE INDIRECT RETRIEVAL OF UNPUBLISHED TECHNICAL MATERIAL
R. L. Patrick
P-2616, August 1962, 4 pp.

The author suggests that ASTIA search its Field of Interest Register if its bibliographic files fail to produce a stipulated minimum number of items in response to a given request.

7. THE SOVIET CLASSIFICATION SCHEME FOR LITERATURE
F. J. Krieger
RM-3325-PR, September 1962, 19 pp.
AD 286 021

Technical and nontechnical literature are both rigorously compartmentalized in the Soviet Union. Krieger compares the national bibliographic schemes of the United States and the USSR.

8. AUTOMATIC PARSING AND FACT RETRIEVAL: A COMMENT ON GRAMMAR, PARAPHRASE, AND MEANING
J. J. Robinson
RM-4005-PR, February 1964, 51 pp.
AD 432 036

The aim of automated parsing is to enable a computer to manipulate information received in the form of natural English sentences. Through the refinement of parsing grammars, insights are to be gained into the nature of meaning, as exemplified, for instance, in the creation of a paraphrase. These insights can then be applied to the development of fact retrieval systems which receive requests for information and output responses in natural English.

9. MECHANIZED DOCUMENTATION: THE LOGIC BEHIND A PROBABILISTIC INTERPRETATION
M. E. Maron
P-2898, April 1964
AD 437 781
In: M. E. Stevens, V. E. Giuliano, L. B. Heilprin, eds., Statistical Association Methods for Mechanized Documentation (National Bureau of Standards Miscellaneous Publication 269), December 15, 1965, pp. 9-13.

This paper discusses the problem of information storage and retrieval in a library. The concepts of probability are required to frame the logic of the problem properly, since the transition from a user's request to the resulting retrieved documents must be schematized as an inverse probability inference. Statistical association techniques are required, for the library computer must be designed to use all the clues and inference techniques available to it.

10. THE LOGIC OF INTERROGATING A DIGITAL COMPUTER
M. E. Maron
P-3006, November 1964, 23 pp.
AD 608 369
In: D. Laszlo, ed., Approaches to Language Data Processing, Mouton & Co., The Hague, 1966 (in press).

An attempt to clarify the problem of how a computing machine must be organized to deal with language in order to respond to interrogation as if it understood meanings and relevance. The Paper is divided into three parts: (1) an outline of the origins and scope of the information sciences and the impetus they have given to this search; (2) a look at two subproblems, literature searching and data retrieval, to suggest how these problems should be framed; and (3) a consideration of the question of comprehension with the aim of specifying how some aspects of knowing can be discussed in a mechanical way and related to the information-flow organization required to generate comprehension-like behavior.

11. RELATIONAL DATA FILE: A TOOL FOR MECHANIZED INFERENCE EXECUTION AND DATA RETRIEVAL
R. Levien and M. E. Maron
RM-4793-PR, December 1965, 89 pp.
AD 625 409

Levien and Maron describe the background and status of a RAND project dealing with automatic data storage and retrieval. Their research emphasizes the development and testing of logical techniques for data retrieval and inference-making. These techniques are being implemented in the form of computer routines, and tested on a large corpus of factual information concerning the field of cybernetics.

A major theoretical decision controlling the designs of a data retrieval system of this sort concerns the nature of the language that will be used. The Levien-Maron system differs from some other question-answering systems in that it does not store information in ordinary or natural language. They discuss their reasons for this decision, examining some of the logical and linguistic aspects of the overall problem, and explaining the selection of an artificial language, an interpreted relational calculus, for use as the information language.

Levien and Maron summarize the theory of relations briefly, showing how relations can be represented in a computer and how some operations on given relations can be computed so as to derive new relations.

They illustrate the operation of their system with some typical data retrieval requests and how they can be executed once incoming data have been mapped into the artificial information language. It also mentions some of the statistical operations, such as correlations and trend analyses, that can be executed.

The key problems of inference are the different kinds of inference that can be made, what role they play in a data retrieval system, and what problems must be solved in order to mechanize inference-making. A key feature of the system is the ability of the user to

frame an inference schema and instruct the machine to execute it. This technique allows the man and the machine to collaborate in deriving a conclusion required to answer a given request for information.

A major part of the Memorandum is devoted to techniques for practical realization of the data file. Input (acquisition, organization, and preprocessing of data before it enters the computer) is a major problem. Resolutions of input problems of data sources, acquisition, extraction, etc. are outlined. Input aids extract data from source documents and enter them onto forms according to precise rules. The computer then translates from forms to relational language sentences. This section discusses the motivation and structure of the input forms.

Output problems are treated against the background of sources and types of output requests that the file will be expected to satisfy. The problem of communicating those requests to the file, and procedures for processing the requests, are discussed.

Techniques for storage and processing are discussed; e.g., how data will be structured in memory, the organization of dictionaries and thesauri, and the steps in the programming process.

The Memorandum concludes with a consideration of the data file as a whole, examining its principal features, application, possible extension, and relation to other data retrieval research. The authors review the question of literature searching and show how the data retrieval system embodies features that permit very effective literature searching. They summarize the hypothesis that context data of the type this system is designed to handle offer a promising approach to more effective literature searching. They present some of the next steps that will be undertaken to extend the capability of the system, and finally summarize and evaluate their work and contrast it with other activities in this field.

12. THE TRANSFORMATION OF SENTENCES FOR INFORMATION RETRIEVAL
J. J. Robinson
P-3243, December 1965, 13 pp.
AD 624 890

Robinson discusses some of the difficulties that must be overcome before automatic retrieval of information expressed in natural language text becomes practical. The major linguistic task is to provide detailed, analytic, recognition grammars with transformational components adequate to deal with the complexities of surface structures of natural sentences.

13. RELATIONAL DATA FILE I: DESIGN PHILOSOPHY
M. E. Maron
P-3408, July 1966, 28 pp.
AD 636 299

Maron's paper and Levien's (sec 7-14) are descriptions of the Relational Data File presented at the Third Annual National Colloquium on Information Retrieval, 1966. Some of the key features of the design philosophy of the system are: information is stored in the form of millions of sentences, processed as necessary to answer questions; all information is uniformly represented in binary relational sentences in a formalized information language; the system is designed to interpret and execute inferences when the answer to a query is merely implicit in the stored data; since literature searching can be accomplished when some of the stored data is about information, sophisticated searches can be conducted through inference techniques dealing with context data.

14. RELATIONAL DATA FILE II: IMPLEMENTATION
R. E. Levien
P-3411, July 1966, 27 pp.
AD 636 311

Related to Maron's (see 7-13), Levien's paper describes the basic features of the implementation of the system: Data are entered by means of special input forms; programs written in FOREMAN, a special form retrieval and manipulation language, derive binary relational sentences from the input tape; a computer code dictionary encodes each sentence as four computer words and full or partial copies of the sentences are entered into four separate files on disk; only subsections of each file are transferred between disk and core; the language at the user's disposal has elementary commands corresponding to the basic operations of arithmetic, set theory and logic, allowing the user to implement simple retrieval, compute counts, correlations and trends, and also execute complex inference schemes.

15. A COMPUTER SYSTEM FOR INFERENCE EXECUTION AND DATA RETRIEVAL
R. E. Levien and M. E. Maron
RM-5085-PR, September 1966, 35 pp.
AD 642 120

This Memorandum is an up-to-date brief and informal description of Relational Data File (see 7-11, 7-13 and 7-14) and the problems that arise in the implementation of such a logical analysis system and how they have been resolved in the now partly operational Relational Data File. Among the specific problems that must be solved are logical and linguistic problems of representing the basic data, file problems of organizing the data for rapid and direct access, programming problems of designing a language for file interrogation, and hardware problems of providing sufficiently large, rapid access storage.

In the Relational Data File, facts are stored in the form of sentences in an artificial information language. Each sentence expresses a binary relationship between two entities. The sentences are directly accessible by content. An intensional file that stores facts about relationships permits many data sentences to be derived when needed rather than stored explicitly. The file of sentences is organized in quadruplicate in a disk memory so that access time is minimized. A programming language has been designed that enables a user to express, in addition to direct retrieval requests, inferential processes that are required to derive implied conclusions. The system is designed for use on an IBM 7044 with a 1301 disk, which holds 232,000 sentences in the information language.

16. A FORMAL SYSTEM FOR THE LOGICAL ANALYSIS OF TEMPORAL RELATIONSHIPS BETWEEN INTERVALS OF TIME
R. Mattison
RM-5279-PR, April 1967, 44 pp.
The system is an applied first-order predicate logic rich enough to express all the usual ordering relations between intervals. The formal language is a first-order language with one binary predicate (interpreted as the "wholly before" that holds between intervals) and two unary operations (interpreted as "the beginning of" and "the ending of" applied to intervals). The semantic completeness of the logical axioms is presupposed, while the syntactic completeness of the temporal axioms is proved. The author also gives a mathematical model of intervals of time which is also a model of the temporal axioms. This model and the temporal axioms are adapted to the needs of the RAND cybernetics data-research project, but philosophically more satisfying axioms are given in an appendix.

17. A SIMPLE SCHEME FOR FORMALIZING DATA RETRIEVAL REQUESTS
F. M. Tonge
RM-5150-PR, May 1967, 40 pp.
The user of the cybernetics data file may be ignorant of computer jargon or the details of file organization. If he is to present data retrieval requests in fairly "natural" language, those requests must be translated. Tonge gives a simple procedure using rules that depend upon the position of words in the request and simple semantic information about words. It presents its translation to the user for modification or correction prior to file search. Tonge gives the rules and several illustrations, and mentions some deficiencies of the present scheme.

8. Content Analysis

1. AUTOMATIC LANGUAGE-DATA PROCESSING IN SOCIOLOGY
D. G. Hays
P-1866, December 1959, 33 pp.
Basic processes of ALDP are described briefly. Techniques based on Osgood's "evaluative assertion analysis" are outlined. The need for research on the relations between linguistic and social variables is emphasized.

2. AUTOMATIC CONTENT ANALYSIS: SOME ENTRIES FOR A TRANSFORMATION CATALOG
D. G. Hays
P-1962, April 1960, 21 pp.
Automatic content analysis would be possible if three steps could be performed automatically: (i) Sentence-structure determination, (ii) structure simplification and regularization, (iii) derivation of sociological observations from the simplified and regularized text. Hays describes the RAND approach to (i) very briefly; step (ii) is discussed at greater length. The third step is examined in terms of Osgood's "evaluative assertion analysis," and some shortcomings of the method, notably its shallowness, are noted.

3. PROCESSING NATURAL LANGUAGE TEXT
D. G. Hays
P-3461, October 1966, 7 pp.
AD 640 658
Two applications of the computer to natural language processing are discussed: for publication of scientific findings and for content analysis by social and behavioral scientists. A documentation system is outlined that starts at the author's typewriter and uses the computer as an editorial aide and as final typist; to help with typesetting; and to produce lists, bibliographies, files, etc. These tasks require little linguistics and are already feasible, although interchange of results will require adopting a standard procedure. Content analysis, on the other hand, presupposes sophisticated linguistic analysis, especially in semantics, which is still lacking. The present successful applications in documentation promise to support further work in linguistics. (Prepared for a Seminar on Computational Linguistics at NIH, October 1966).

9. Psycholinguistics

1. AN INFORMATION PROCESSING THEORY OF VERBAL LEARNING
E. A. Feigenbaum
P-1817, October 1959, 169 pp.
A theory of some elementary forms of human symbolic learning: memorization, discrimination, association, and attention direction. The theory is concerned with mental activity at the level of the

processing of information symbols, which are the basic units manipulated. The precise statement of the theory is given in the language of a digital computer, specifically as a set of programs in IPL-V called EPAM (Elementary Perceiver and Memorizer). The author deals generally with information structures and processes for discrimination and association learning, and specifically with behavior in the standard rote-learning task. The behavior of the model is compared with that of human subjects.

2. THE SIMULATION OF VERBAL LEARNING BEHAVIOR
E. A. Feigenbaum
P-2235, March 1961, 30 pp.
Proceedings of the Western Joint Computer Conference, vol. 19 (1961), pp. 129-132.

An abbreviated description of EPAM and its behavior.

3. PERFORMANCE OF A READING TASK BY AN ELEMENTARY PERCEIVING AND MEMORIZING PROGRAM
E. A. Feigenbaum and H. A. Simon
P-2358, July 1961, 14 pp.
Behavioral Science, vol. 8, no. 1 (January 1963), pp. 72-76.

Experiments show that the mechanisms postulated in EPAM for rote memory tasks are adequate for simulating, at least macroscopically, the processes used by human beings in acquiring the ability to read and understand printed words. The authors provide a summary description of the EPAM program, mention the main processes it uses in rote memory tasks, and describe how these processes are used in learning to read.

4. A THEORY OF THE SERIAL POSITION EFFECT
E. A. Feigenbaum and H. A. Simon
P-2375, July 1961, 35 pp.

The theory presented is sufficient to predict (qualitatively and quantitatively) the shape of the serial error curve. In addition, other rote-learning phenomena are explained. The theory postulates a serial information processing mechanism that learns (on the average) one item from a serial list every k seconds, has a very small immediate memory span, and uses an anchor-point processing strategy for organizing its learning effort over time. Two ways described to make predictions from the postulates are by a computer programmed to process information and by a simple mathematical model.

5. A NET TO SIMULATE MORSE-CODE LEARNING
I. Barr
RM-3850-PR, January 1964, 33 pp.
AD 429 107

The design of a neural net which can learn and recall six letters of Morse Code is presented. The net, called MCN, can learn these coded equivalents in any order. By cascading the MCN with a sequence-recall net, a new net is created, called T, which is able to encode words after being taught the individual letter codes. The method of combining these two nets and a modification of the referenced sequence-recall net are also discussed.

6. STUDIES IN INFORMATION PROCESSING THEORY: SIMILARITY AND FAMILIARITY IN VERBAL LEARNING
H. A. Simon and E. A. Feigenbaum
RM-3979-PR, February 1964, 39 pp.
AD 430 739

A presentation of results obtained by simulating various verbal learning experiments with the Elementary Perceiving and Memorizing Program (EPAM), an information processing theory of verbal learning. Predictions were generated for experiments manipulating intra-list similarity (Underwood), inter-list similarity (Bruce), and familiarity and meaningfulness. The stimulus materials were nonsense syllables, learned in paired-associate fashion. The predictions made by the model are generally in good agreement with the experimental data. The fit of the EPAM predictions to the Chenzoff data is particularly significant since it lends support to the hypothesis that the mechanism by which a high degree of meaningfulness of items facilitates learning is the high familiarity of these items.

7. CONSTRUCTION OF A SIMULATION PROCESS FOR INITIAL PSYCHIATRIC INTERVIEWING
N. L. Gilbreath, R. E. Bellman, M. B. Friend, and L. Kurland
P-2933, June 1964, 13 pp.
AD 602 976

A description of the construction of a simulation of an initial psychiatric interview, regarded as an example of an adaptive, multi-stage decision process.

8. ON THE CONSTRUCTION OF A SIMULATION OF THE INITIAL PSYCHIATRIC INTERVIEW
R. E. Bellman, M. B. Friend, and L. Kurland
RM-4044-PR, July 1964, 58 pp.
(National Institute of Health)
AD 602 649

A description of the construction of a simulation of an initial psychiatric interview. It can be regarded as an example of an adaptive, multi-stage decision process.

9. COMPUTER SIMULATION OF HUMAN BEHAVIOR
E. A. Feigenbaum
P-2905, May 1964, 14 pp.
AD 601 075

This paper discusses computer simulation of human behavior with particular emphasis on human cognitive behavior (learning and problem-solving processes). Included are remarks about the simulator of this behavior, the digital computer.

10. GENERALIZATION OF AN ELEMENTARY PERCEIVING AND MEMORIZING MACHINE
E. A. Feigenbaum and H. A. Simon
P-2555, March 1962, 16 pp.
Proceedings of IFIP Congress 62, North Holland, 1963.

An exploration of the Elementary Perceiver and Memorizer (EPAM) III model, which is concerned with simulating the behavior observed in psychological experiments on meaningfulness in verbal learning. In this model, EPAM information processes and structures are generalized to deal with stimulus objects of arbitrary complexity. Discrimination processes discriminate objects on the basis of properties of the objects themselves or on the basis of properties of constituent subobjects. The learning processes of the EPAM III model provide an associative mechanism by means of which earlier learning is brought to bear in a useful way on later learning.

11. LINGUISTIC RELATIVITY AND THE LANGUAGE LEARNING PROCESS
R. M. Schwarcz
RM-5210-PR, December 1966, 13 pp.
AD 644 929

A five-stage analysis of the language-learning process, and an investigation of whether this analysis supports or contradicts the Whorfian hypothesis of linguistic relativity. The syntactic constructions of a language influence the types of conceptual relationship that the child perceives. Words and phrases grouped together by the conventions of the language have a determining effect on conceptual representations associated with them. This effect is somewhat mitigated by the transformational phase of language learning, which opens up the full range of stylistic devices available in a language, thus extending the number of percepts and concepts that may gain expression in that language. However, earlier perceptual and conceptual habits will probably persist throughout a person's adult life unless he is forced to change by some new experience, such as the mastery of a different language.

12. STEPS TOWARD A MODEL OF LINGUISTIC PERFORMANCE: A PRELIMINARY SKETCH
R. M. Schwarcz
RM-5214-PR, January 1967, 43 pp.
AD 646 598

The model of linguistic performance outlined here is based on the author's theoretical approach to linguistics described in 9-11. The task of formulating a model of linguistic performance is discussed and an approach to embodying the model as a computer program is proposed.

The methodology of current linguistic theory is criticized for several of its features that render it inapplicable to a realistic model of performance, and remedies for these deficiencies are proposed. The syntactic and conceptual data structures, inference rules, generation and understanding mechanisms, and learning mechanisms proposed for the model are all described. The learning process is formulated as a series of five stages, and the roles of nonlinguistic feedback and inductive generalization relative to these stages are described. Finally, the implications of a successful performance model for linguistic theory, linguistic applications of computers, and psychological theory are discussed.

10. Character Readers

1. A DIGITAL SIMULATION OF AN AIDED ADAPTIVE CHARACTER READING MACHINE
P. Baran and G. Estrin
P-1989, May 1960, 26 pp.

The pattern-recognition system here simulated on a digital computer uses an initial man-machine learning phase. Transformations on a deformed set of 48 samples of each of ten numerals are used to form separation filters, while a second set of 480 similarly varied numerals serve as the "unknown" characters that are examined. Measured probability density distributions of the inked areas of all characters are established, and a weighted stencil or filter is created to distinguish each character relative to the possible set of characters. This experiment demonstrates the extent to which the actual value of the best "score of match" between the unknown and each character in the set provides confidence in recognition. Whenever the best score is too low, it is possible to call for more complex processes to aid recognition, permitting the construction of recognition systems of greater accuracy than the basic reading mechanism.

2. AN AIDED ADAPTIVE CHARACTER READER FOR MACHINE TRANSLATION OF LANGUAGES
P. Baran and G. Estrin
P-1990, May 1960, 45 pp.

Here the authors show that the elementary model of P-1989 may aid in constructing a fast-input device for a language translation machine if it was able to make use of frequency distribution characteristics of the dictionary. A possible implementation with a raw character reading rate up to 500 characters a second appears feasible.

3. THE RAND TABLET: A MAN-MACHINE GRAPHICAL COMMUNICATION DEVICE
M. R. Davis and T. O. Ellis
RM-4122-ARPA, August 1964, 28 pp.
AD 444 103
(Advanced Research Project Agency)
Proceedings of the Fall Joint Computer Conference, Spartan Books, Baltimore, 1964, pp. 325-331.

This Memorandum describes a low-cost, two-dimensional graphic input tablet and stylus developed at the RAND Corporation for conducting research on man-machine graphic communications. The tablet is a printed-circuit screen complete with printed-circuit capacitative-coupled encoders with only 40 external

connections. The writing surface is a 10" x 10" area with a resolution of 100 lines per inch in both x and y. Thus, it is capable of digitizing 10^6 discrete locations with excellent linearity, allowing the user to "write" in a natural manner. The system does not require a computer-controlled scanning system to locate and track the stylus. Several institutions have recently installed copies of the tablet in research environments. It has been in use at RAND since September 1963.

4. ABSTRACTION AND PATTERN CLASSIFICATION
R. E. Bellman, R. E. Kalaba, and L. A. Zadeh
RM-4307-PR, October 1964, 20 pp.
AD 607 500

A preliminary discussion of a general framework for the treatment of pattern-recognition problems. The authors define "fuzzy" set, then show how this may be used in a sequential experimental procedure to ascertain whether a symbol is a member of a particular set or not.

5. COMPUTER RECOGNITION OF ON-LINE, HANDWRITTEN CHARACTERS
M. I. Bernstein
RM-3753-ARPA, October 1964, 33 pp.
AD 451 231
(Advanced Research Project Agency)

Discussion of a method for recognizing single, hand-written characters using an on-line graphical input device such as a digitizing pantograph, light pen, or RAND Graphic Input Tablet, as the primary information source. Basically, the method consists of filtering and smoothing the input stream to eliminate as much redundancy as possible. Direction of the stylus movement is quantized into one of eight directions, allowing each stroke of a character to be described as a series of connected straight-line segments. By eliminating various measures on the stroke, the description is size-, position-, and rotation-independent. To restore some rotational orientation and to discriminate between open, closed, and multi-stroke characters, end-point comparisons are added to the description.

6. ON THE DEVELOPMENT OF EQUITABLE GRAPHIC I/O
T. O. Ellis and W. L. Sibley
P-3415, July 1966, 13 pp.
AD 637 781

The desire for direct interaction between man and machine has led to the study of computer interpretation of free-hand motions of a stylus and the real-time responses to these motions. The operating environment of GRAIL—Graphic Input Language—under development at RAND is discussed. The system utilizes elements of pictorial and verbal languages.

11. Translations

1. EXACT METHODS IN LINGUISTIC RESEARCH
O. S. Akhmanova, I. A. Mel'chuk, R. M. Frumkina, E. V. Paducheva
(Translated by D. G. Hays and D. V. Mohr)
R-397, September 1963
University of California Press, Berkeley, 1963
AD 420 565

An attempt to illuminate several scientific results in the application of exact methods to linguistic research. Often termed "mathematical linguistics," the essence of this new approach and its content consists not of creating some special kind of "linguistics," but rather of perfecting, of making accurate, reliable, and modern the methods of conventional linguistic research. This Report deals with: (a) questions of general linguistics which relate the following discussions of concrete methods of exact study and description of linguistic phenomena to earlier linguistics; (b) the place and role of machine translation in contemporary linguistics in both a theoretical and a practical sense; (c) possible applications of statistical methods to linguistic research, together with a discussion of basic principles and concepts of statistical analysis; (d) possible applications of information theory to language study. The Report does not treat linguistic applications of "non-quantitative" mathematics--in particular, mathematical logic. This work was translated in order to show the direction and scope of the best Russian research in modern linguistics.

2. TWO OPERATORS FOR DETERMINING AGREEMENT FOR AUTOMATIC SYNTACTIC ANALYSIS
I. A. Mel'chuk
(Translated by D. V. Mohr)
RM-3190, June 1962

Two interrelated programs are presented for the automatic analysis of Russian sentences in any language characterized by grammatic agreement among certain word classes that must be tested to derive the proper syntactic structure for each sentence being processed.

3. PROBLEMS OF ALGORITHMIC COMPOSITION OF SUBJECT INDEXES (Brief Survey of Foreign Literature)
B. V. Iakushin
Scientific-Technical Information (Nauchno-Tekhnicheskaya Informatsiya)
(Translated by J. B. Gazley and W. B. Holland
LT-65-102, March 15, 1966, pp. 22-25

Major trends in developing methods of machine assignment of subject indexes are examined. Two approaches are discernible: a) the use of subject descriptors based on key words extracted from the document (usually from the title); b) the use of predetermined subject categories under which documents are indexed. It is noted that the systems that rely on statistical characteristics of the appearance of key words in the documents are the more promising.

DEPOSIT LIBRARIES IN THE UNITED STATES

Alabama
- Auburn: Auburn University*

Arizona
- Tempe: Arizona State University*
- Tucson: University of Arizona*

California
- Berkeley: University of California
- Goleta: General Motors Corporation*
- Irvine: University of California*
- Los Angeles: University of California
  University of Southern California*
- Pasadena: California Institute of Technology
- Riverside: University of California*
- Stanford: Stanford University

Colorado
- Denver: Denver Public Library

Connecticut
- New Haven: Yale University

Washington, D.C.
- The Library of Congress

Florida
- Gainesville: University of Florida

Georgia
- Atlanta: Georgia Institute of Technology

Hawaii
- Honolulu: University of Hawaii (partial deposit, beginning with R-287, RM-1606, and P-787)

Illinois
- Chicago: University of Chicago
- Des Plaines: Borg-Warner Corporation*
- East St. Louis: Parks College of Aeronautical Technology*
- Urbana: University of Illinois

Indiana
- Lafayette: Purdue University

Iowa
- Ames: Iowa State University of Science and Technology
- Iowa City: State University of Iowa

---

*Subscription library; access to collection may be restricted.

Kansas
Manhattan: Kansas State University*

Kentucky
Lexington: University of Kentucky

Louisiana
New Orleans: Tulane University*

Maryland
Baltimore: The Johns Hopkins University
College Park: University of Maryland*

Massachusetts
Amherst: University of Massachusetts*
Cambridge: Harvard University
Massachusetts Institute of Technology
Chestnut Hill: Boston College*

Michigan
Ann Arbor: University of Michigan
Detroit: Detroit Public Library*
East Lansing: Michigan State University*
Kalamazoo: Western Michigan University*

Minnesota
Minneapolis: University of Minnesota

Missouri
Columbia: University of Missouri*
Kansas City: Linda Hall Library
St. Louis: Washington University

Nebraska
Lincoln: The University of Nebraska

Nevada
Las Vegas: Nevada Southern University*

New Jersey
Princeton: Princeton University

New York
Buffalo: State University of New York*
Ithaca: Cornell University
New York: Columbia University
The New York Public Library
Stony Brook, L. I.: State University of New York*
Syracuse: Syracuse University*

North Carolina
Durham: Duke University

---

*Subscription library; access to collection may be restricted.

Ohio
Cleveland: Case Institute of Technology*
Cleveland Public Library
Columbus: Ohio State University

Oklahoma
Norman: The University of Oklahoma

Oregon
Corvallis: Oregon State University

Pennsylvania
Pittsburgh: Carnegie Library of Pittsburgh
University Park: Pennsylvania State University*

Rhode Island
Providence: Brown University

Tennessee
Memphis: Memphis State University*
Nashville: Joint University Libraries

Texas
Austin: The University of Texas
College Station: Texas A & M University*
Houston: Rice University

Utah
Provo: Brigham Young University*
Salt Lake City: University of Utah

Virginia
Blacksburg: Virginia Polytechnic Institute*
Charlottesville: University of Virginia

Washington
Renton: The Boeing Company*
Seattle: University of Washington

Wisconsin
Madison: University of Wisconsin

Wyoming
Cheyenne: Wyoming State Library*

---

*Subscription library; access to collection may be restricted.

DEPOSIT LIBRARIES ABROAD

Australia
- Sydney: The University of Sydney (partial deposit beginning with R-287, RM-1606, and P-787)

Canada
- Ottawa: Defence Research Board*
  National Research Council of Canada

England
- Boston Spa, Yorkshire: National Lending Library for Science and Technology

France
- Paris: S.E.M.A. Documentation*
  Société Francaise de Recherche Opérationnelle

Germany
- Bochum-Querenburg: Universitätsbibliothek der Ruhruniversität*
- Ebenhausen/Isartal: Stiftung Wissenschaft und Politik*
- Frankfurt a.M: Stadt- und Universitäts-Bibliothek
- Munich: Industrieanlagen-Betriebsgesellschaft Einsteinstrasse*

India
- Kanpur: Indian Institute of Technology*
- New Delhi: Defence Science Laboratory*
  Indian Council of World Affairs*
  Indian Institute of Public Administration*

Israel
- Haifa: Technion - Israel Institute of Technology*
- Tel-Aviv: Mamram Computer Center*

Japan
- Tokyo: National Diet Library (complete deposit except for R-245)

Puerto Rico
- Rio Piedras: Universidad de Puerto Rico*

Sweden
- Nykoping: Akiebolaget Atomenergi

---

* Subscription library; access to collection may be restricted.

Author Index

PUBLICATION NUMBER INDEX

RAND MEMORANDA

RAND MEMORANDA (CONT.)

5345	1-23	1989	10-1	3006	7-10
		1990	10-2	3069	1-18
PAPERS		2142	1-1	3101	4.2-10
187	2.2-1	2197	4.2-1	3112	4.3-2
1218	5.3-1	2206	5.2-4	3136	4.2-11
1241	5.1-2	2235	9-2	3242-1	2.2-6
1251	3-3	2279	7-2	3243	7-12
1321	1-9	2315	2.1-3	3244	1-21
1328	1-8	2327	1-11	3408	7-13
1588	1-6	2339	2.2-2	3411	7-14
1624	3-5	2349	4.2-7	3413	2.2-7
1632	5.2-1	2358	9-3	3415	10-6
1720	5.2-2	2375	9-4	3436	1-22
1771	4.2-6	2455	7-3	3439	2.2-8
1817	9-1	2471	7-4	3461	8-3
1866	8-2	2526	4.3-1	3476	4.2-16
1896	1-12	2555	9-10	3497	4.1-6
1900	1-7	2586	1-2	3500	3-10
1909	4.2-3	2599	3-6		
1910	2.1-1	2616	7-5	REPORT	
1926	4.2-2	2879	1-13	397	11-1
1941	5.3-4	2892	1-14		
1962	8-3	2898	7-9		
1978	5.2-3	2905	9-9		
1984	2.1-2	2933	9-7		