

**UNCLASSIFIED**

**AD** 435103

**DEFENSE DOCUMENTATION CENTER**

**FOR**

**SCIENTIFIC AND TECHNICAL INFORMATION**

**CAMERON STATION, ALEXANDRIA, VIRGINIA**



**UNCLASSIFIED**

**NOTICE:** When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

435103

LRC 64-P19

435103

CATALOGED BY DDC

AS AD No. \_\_\_\_\_

LINGUISTICS RESEARCH CENTER The University of Texas

Report No. 19

THIRD QUARTERLY PROGRESS REPORT | NOVEMBER 1963—31 JANUARY 1964

U. S. ARMY ELECTRONICS RESEARCH AND DEVELOPMENT LABORATORY MACHINE LANGUAGE TRANSLATION STUDY

CONTRACT NUMBER DA 36-039 AMC-02162(E) FILE NUMBER 5704-PM-63-91 (6514)

NO OTS

APR 16 1964  
TISA B

QUALIFIED REQUESTORS MAY OBTAIN  
COPIES OF THIS REPORT FROM ASTIA

ASTIA RELEASE TO OTS IS  
NOT AUTHORIZED.

<p>The University of Texas MACHINE LANGUAGE TRANSLATION STUDY W. P. Lehmann Chief Investigator</p> <p>Report No. 19, Third Quarterly Progress Report, 1 November 1963 to 31 January 1964, 38 pp. Signal Corps Contract DA 36-039 AMC-02162 (E) File No. 5704-PM-63-91 (6514).</p> <p>Progress is reported in the development of two generalized computer systems: the first designed to implement automatic translation of languages, the second to support basic research in linguistics. The systems complement each other in that programs prepared for each are applicable to, and needed in, the other. The common system design contains three sections: one for control, a second for language data processing, and a third for linguistic information processing. The first two sections are now operational. In the third, Monolingual Recognition programs for performing lexical and syntactic analysis and display have been made operational. The programs, which have been converted to operate on the IBM 7040 as well as the IBM 7090 computer, are being tested with English, German, Russian, and Chinese language data.</p>	<p>Unclassified</p> <p>Machine Language Translation Study</p> <p>1. Signal Corps Contract DA 36-039 AMC-02162(E)</p>	<p>The University of Texas MACHINE LANGUAGE TRANSLATION STUDY W. P. Lehmann Chief Investigator</p> <p>Report No. 19, Third Quarterly Progress Report, 1 November 1963 to 31 January 1964, 38 pp. Signal Corps Contract DA 36-039 AMC-02162 (E) File No. 5704-PM-63-91 (6514).</p> <p>Progress is reported in the development of two generalized computer systems: the first designed to implement automatic translation of languages, the second to support basic research in linguistics. The systems complement each other in that programs prepared for each are applicable to, and needed in, the other. The common system design contains three sections: one for control, a second for language data processing, and a third for linguistic information processing. The first two sections are now operational. In the third, Monolingual Recognition programs for performing lexical and syntactic analysis and display have been made operational. The programs, which have been converted to operate on the IBM 7040 as well as the IBM 7090 computer, are being tested with English, German, Russian, and Chinese language data.</p>	<p>Unclassified</p> <p>Machine Language Translation Study</p> <p>1. Signal Corps Contract DA 36-039 AMC-02162(E)</p>
<p>The University of Texas MACHINE LANGUAGE TRANSLATION STUDY W. P. Lehmann Chief Investigator</p> <p>Report No. 19, Third Quarterly Progress Report, 1 November 1963 to 31 January 1964, 38 pp. Signal Corps Contract DA 36-039 AMC-02162 (E) File No. 5704-PM-63-91 (6514).</p> <p>Progress is reported in the development of two generalized computer systems: the first designed to implement automatic translation of languages, the second to support basic research in linguistics. The systems complement each other in that programs prepared for each are applicable to, and needed in, the other. The common system design contains three sections: one for control, a second for language data processing, and a third for linguistic information processing. The first two sections are now operational. In the third, Monolingual Recognition programs for performing lexical and syntactic analysis and display have been made operational. The programs, which have been converted to operate on the IBM 7040 as well as the IBM 7090 computer, are being tested with English, German, Russian, and Chinese language data.</p>	<p>Unclassified</p> <p>Machine Language Translation Study</p> <p>1. Signal Corps Contract DA 36-039 AMC-02162(E)</p>	<p>The University of Texas MACHINE LANGUAGE TRANSLATION STUDY W. P. Lehmann Chief Investigator</p> <p>Report No. 19, Third Quarterly Progress Report, 1 November 1963 to 31 January 1964, 38 pp. Signal Corps Contract DA 36-039 AMC-02162 (E) File No. 5704-PM-63-91 (6514).</p> <p>Progress is reported in the development of two generalized computer systems: the first designed to implement automatic translation of languages, the second to support basic research in linguistics. The systems complement each other in that programs prepared for each are applicable to, and needed in, the other. The common system design contains three sections: one for control, a second for language data processing, and a third for linguistic information processing. The first two sections are now operational. In the third, Monolingual Recognition programs for performing lexical and syntactic analysis and display have been made operational. The programs, which have been converted to operate on the IBM 7040 as well as the IBM 7090 computer, are being tested with English, German, Russian, and Chinese language data.</p>	<p>Unclassified</p> <p>Machine Language Translation Study</p> <p>1. Signal Corps Contract DA 36-039 AMC-02162(E)</p>

<p>The University of Texas MACHINE LANGUAGE TRANSLATION STUDY W. P. Lehmann Chief Investigator</p> <p>Report No. 19, Third Quarterly Progress Report, 1 November 1963 to 31 January 1964, 38 pp. Signal Corps Contract DA 36-039 AMC-02162 (E) File No. 5704-PM-63-91 (6514).</p> <p>Progress is reported in the development of two generalized computer systems: the first designed to implement automatic translation of languages, the second to support basic research in linguistics. The systems complement each other in that programs prepared for each are applicable to, and needed in, the other. The common system design contains three sections: one for control, a second for language data processing, and a third for linguistic information processing. The first two sections are now operational. In the third, Monolingual Recognition programs for performing lexical and syntactic analysis and display have been made operational. The programs, which have been converted to operate on the IBM 7040 as well as the IBM 7090 computer, are being tested with English, German, Russian, and Chinese language data.</p>	<p>Unclassified</p> <p>1. Machine Language Translation Study Contract DA 36-039 AMC-02162(E)</p>	<p>The University of Texas MACHINE LANGUAGE TRANSLATION STUDY W. P. Lehmann Chief Investigator</p> <p>Report No. 19, Third Quarterly Progress Report, 1 November 1963 to 31 January 1964, 38 pp. Signal Corps Contract DA 36-039 AMC-02162 (E) File No. 5704-PM-63-91 (6514).</p> <p>Progress is reported in the development of two generalized computer systems: the first designed to implement automatic translation of languages, the second to support basic research in linguistics. The systems complement each other in that programs prepared for each are applicable to, and needed in, the other. The common system design contains three sections: one for control, a second for language data processing, and a third for linguistic information processing. The first two sections are now operational. In the third, Monolingual Recognition programs for performing lexical and syntactic analysis and display have been made operational. The programs, which have been converted to operate on the IBM 7040 as well as the IBM 7090 computer, are being tested with English, German, Russian, and Chinese language data.</p>	<p>Unclassified</p> <p>1. Machine Language Translation Study Contract DA 36-039 AMC-02162(E)</p>
<p>The University of Texas MACHINE LANGUAGE TRANSLATION STUDY W. P. Lehmann Chief Investigator</p> <p>Report No. 19, Third Quarterly Progress Report, 1 November 1963 to 31 January 1964, 38 pp. Signal Corps Contract DA 36-039 AMC-02162 (E) File No. 5704-PM-63-91 (6514).</p> <p>Progress is reported in the development of two generalized computer systems: the first designed to implement automatic translation of languages, the second to support basic research in linguistics. The systems complement each other in that programs prepared for each are applicable to, and needed in, the other. The common system design contains three sections: one for control, a second for language data processing, and a third for linguistic information processing. The first two sections are now operational. In the third, Monolingual Recognition programs for performing lexical and syntactic analysis and display have been made operational. The programs, which have been converted to operate on the IBM 7040 as well as the IBM 7090 computer, are being tested with English, German, Russian, and Chinese language data.</p>	<p>Unclassified</p> <p>1. Machine Language Translation Study Contract DA 36-039 AMC-02162(E)</p>	<p>The University of Texas MACHINE LANGUAGE TRANSLATION STUDY W. P. Lehmann Chief Investigator</p> <p>Report No. 19, Third Quarterly Progress Report, 1 November 1963 to 31 January 1964, 38 pp. Signal Corps Contract DA 36-039 AMC-02162 (E) File No. 5704-PM-63-91 (6514).</p> <p>Progress is reported in the development of two generalized computer systems: the first designed to implement automatic translation of languages, the second to support basic research in linguistics. The systems complement each other in that programs prepared for each are applicable to, and needed in, the other. The common system design contains three sections: one for control, a second for language data processing, and a third for linguistic information processing. The first two sections are now operational. In the third, Monolingual Recognition programs for performing lexical and syntactic analysis and display have been made operational. The programs, which have been converted to operate on the IBM 7040 as well as the IBM 7090 computer, are being tested with English, German, Russian, and Chinese language data.</p>	<p>Unclassified</p> <p>1. Machine Language Translation Study Contract DA 36-039 AMC-02162 (E)</p>

Report No. 19  
THIRD QUARTERLY PROGRESS REPORT  
1 NOVEMBER 1963 - 31 JANUARY 1964

W. P. Lehmann, Director  
E. D. Pendergraft, Associate Director

Prepared for  
U. S. Electronics Research and Development Laboratory  
Fort Monmouth, New Jersey  
Contract No. DA 36-039 AMC-02162 (E)  
File No. 5704-PM-63-91 (6514)

LINGUISTICS RESEARCH CENTER  
The University of Texas  
Box 7247, University Station  
Austin, Texas 78712

LRC 64-P19

February 1964

## TABLES OF CONTENTS

1	Abstract .....	5
2	Purpose .....	7
3	Publications, Lectures, Reports, and Conferences..	11
4	Research Summary .....	15
	4.1 Foundations .....	15
	4.2 Language Data Collection .....	16
	4.3 Systems Development .....	18
5	Progress in the Quarter .....	19
	5.1 Systems .....	19
	5.1.1 Programming .....	19
	5.1.2 Operations .....	23
	5.2 Descriptive Linguistics .....	23
	5.2.1 English .....	24
	5.2.2 German .....	26
	5.2.3 Russian .....	26
	5.2.4 Chinese .....	27
	5.3 Theoretical Linguistics .....	27
6	Conclusions .....	29
7	Planning for the Next Quarter .....	31
	REFERENCES .....	33
	PERSONNEL .....	35
	TABLES	
5-1	Current Status of System Programs .....	20



1            ABSTRACT

Progress is reported in the development of two generalized computer systems: the first designed to implement automatic translation of languages, the second to support basic research in linguistics. The systems complement each other in that basic programs prepared for each are applicable to, and needed in, the other. The common system design contains three sections: one for control, a second for language data processing, and a third for linguistic information processing. The first two sections are now essentially operational. In the third, Monolingual Recognition programs for performing lexical and syntactic analysis and display have been made operational. The programs, which have been converted to operate on the IBM 7040 as well as the IBM 7090 computer, are being tested with English, German, Russian, and Chinese language data.

The Machine Language Translation Study has been in progress since May 1959 under sponsorship of U.S. Army Electronics Research and Development Laboratory. The project has a long-range but primarily practical purpose: to provide an automatic translation system with sufficient capacity and generality to be useful in a military environment.

The study is especially concerned with translation of foreign languages into English. Although German was selected as the input language to be used in initial testing, the need for translation techniques which would be applicable to other languages was recognized in the research from its inception. Thus, translation of Russian and Chinese are also contemplated.

A long-range research approach was taken to provide the opportunity for investigation of general principles underlying processes occurring in translation, and for subsequent application of the resulting techniques. The work accordingly has two closely related objectives: one predominantly scientific and the other practical in the sense of being directed toward military applications.

The scientific objective will be attacked in two phases. First, a general theory of translation will be

formulated and verified through descriptive research on the languages mentioned above. The goal in this phase will be to describe structural alternatives which are present in the empirical data of each language. Secondly, heuristic processes which make efficient choices among these alternatives will be developed as a basis for useful translation applications.

The computer system incorporating these heuristic features, which will optimize its usefulness in a military environment, will be referred to as the Language Translation System.

A second project, under sponsorship of the National Science Foundation, was initiated in September 1961 when it became apparent that the considerable labor involved in reaching this scientific objective could be expedited by a separate computer system designed specifically to support linguistic research. The study, entitled "Development of a Linguistic Computer System," is essentially concerned with automating processes necessary to translation research rather than the process of translation itself.

The computer system will carry out the translation process, but it will do so for the purpose of displaying the alternatives that are available to processes making heuristic choices. A probability will be computed

for each alternative to further facilitate study of heuristic translation processes. Mechanical assistance in language data collection and processing will also be provided by the system.

The computer system incorporating these supporting features, which will optimize its usefulness as an environment for theoretical and descriptive linguistic research, will be referred to as the Linguistics Research System.

The function of this report is to describe progress toward implementation of the Language Translation System during the period from 1 November to 31 January 1964. Because it will be most convenient to implement this system through an adaptation of certain basic parts of the Linguistics Research System, the two projects are currently cooperating in developing those computer programs which they will use in common. Part of the progress reported below has therefore been accomplished by the second project.

Individual contributions of the two projects have not been distinguished because any attempt to do so, at the present time, would be artificial.

3 PUBLICATIONS, LECTURES, REPORTS AND CONFERENCES

Dr. Tosh presented a lecture on machine translation research to the Modern Language Club of Texas A. & M. University on 11 November.

Dr. Prem Kishore Kulshrestha of the India Institute of Technology, Bombay, India, visited LRC on 18-19 November. He was visiting the United States under UNESCO sponsorship to gain familiarity with machine translation work applicable to Russian-English texts. Dr. Kulshrestha proposes to return and use LRC facilities to prepare a description of Hindi.

Dr. Robert H. Owens of the National Science Foundation visited LRC on 6 December. He discussed research at the Center with Dr. Lehmann and Mr. Pendergraft.

Dr. John Johnston of the University of Kansas visited LRC on 13 December, discussing with Mr. Pendergraft and Dr. Tosh his work in programming languages and LRC descriptive techniques for natural languages.

Mr. Stephen B. Smith of Thompson, Ramo Wooldridge, Inc., visited the Center on 13 December. As Associate Project Manager of the NSF Text Collection Center Study, Mr. Smith was interested in a detailed survey of the facilities and research of LRC.

Mr. Robert Dunn of USAERDL, Fort Monmouth, visited the Center on 19-20 December. He discussed contract details

with Dr. Lehmann and Mr. Pendergraft, and reviewed LRC research with Dr. Tosh, Mr. Jonas, and Mr. Estes.

Mr. Paul Jones of Arthur D. Little, Inc., visited LRC on 19-20 December. Currently working on an associative information retrieval system, he was interested in natural language processing and the automatic classification techniques being investigated at the Center. He conferred with Mr. Pendergraft, Dr. Tosh, Mr. Jonas, and Dr. Dale.

Drs. Lehmann, Tosh, and Joynes and Mr. Pendergraft attended the meeting of the Linguistics Society of America in Chicago on 28-30 December. The following papers were read:

Lehmann: "Vowel Systems, Especially that of PIE"

Tosh: "Development of Automatic Grammars"

Dr. Lehmann visited in Cairo, Egypt, on 1-2 January to discuss projects in machine translation with representatives of the Ministry for Research.

Dr. Lehmann attended the 26th Congress of Orientalists in New Delhi, India, on 3-10 January, and read a paper on linguistic research.

Dr. Richard N. Adams, Department of Anthropology, University of Texas, visited LRC on 6 January. He proposes to use LRC grammar coding techniques to describe behavioral patterns in a Guatemalan sub-culture.

Mr. K. C. Hageman, General Dynamics Corp., Fort Worth, visited LRC on 9 January. Concerned with processing data being generated in development of the F-111 weapons system, Mr. Hageman discussed with Dr. Dale possible applications of clump theory, which is being investigated at LRC.

A working paper on German nouns was completed and distributed during the quarter: "Report on German Noun Coding," G. R. Lewis and L. W. Tosh--LRC 63-WDG1, November 1963.

Three research groups of the Linguistics Research Center are participating in the study: the (1) Theoretical Linguistics group with skills in mathematics and logic, (2) Systems group proficient in computer programming and operations, and (3) Descriptive Linguistics group of linguists specializing in German, English, Russian or Chinese.

#### 4.1 Foundations

Precise theories of linguistic structure are a prerequisite to successful applications in linguistic information processing, whether for mechanical translation or for other processes involving analysis of information content, such as abstracting or information storage and retrieval. A working hypothesis has been formulated which formalizes present objectives in analysis of syntactic and semantic content [1,2]. The hypothesis includes certain types of transformations [3,4] which must be taken into account by generalized algorithms performing syntactic and semantic analysis.

A general theory of translation has been based on this foundation. Within the theory, an interlingual description of relations among lexical, syntactic, or semantic units of various languages provides different



kinds of transfer from one language to another. Thus the general hypothesis underlying programming for the Language Translation System is considered to be completed. Theoretical work is now concerned with mathematical analysis which may lead to increased operational efficiency of the Language Translation System.

An attempt is being made, under the National Science Foundation grant, to extend the hypothesis to the level of pragmatic description. Such an extension is believed to be a necessary step toward information retrieval, artificial intelligence, and automatic abstracting applications.

#### 4.2 Language Data Collection

The enormous task of collecting and verifying structural data for the various languages which would be of interest in linguistic information processing applications must be undertaken systematically on a long-range basis if substantial progress is to be made. The Linguistics Research System is primarily intended to support this important research function. The computer system will be capable of maintaining large stores of language data through accounting procedures and of manipulating the data within linguistic information processing algorithms.

Since such research is best performed within an academic environment, language data collection and verification is being pursued chiefly under the National Science Foundation grant. Resultant data are being made available to the Language Translation System in exchange for algorithms which may be used in the Linguistics Research System.

Descriptive investigations of English and German have been in progress for more than three years. Russian and Chinese studies are of much more recent origin, dating from the fourteenth quarter. Most of the work for these languages has been in the area of syntax, though basic semantic data have been collected through the examination of synonyms and equivalent expressions. German and English syntactic studies have been oriented to a specific corpus taken from Eduard Ruechardt's Sichtbares und Unsichtbares Licht and its English translation [5]. German and English dictionaries have also been based on Der Sprach-Brockhaus and Webster's New Collegiate Dictionary, respectively. Russian and Chinese studies are likewise text-oriented. Three articles from Voprosy Ekonomiki [6] are being used for the former, and a Chinese text on language teaching [7] for the latter.

### 4.3 Systems Development

Programming criteria for translation algorithms based on lexical, syntactic, and semantic transfer have been derived from the working hypothesis. These, together with programs which maintain and compile lexical, syntactic, semantic, and interlingual description, have been described elsewhere [1]. An executive routine, referred to as the Control program, has been completed, as have all programs which maintain and compile language descriptions. Automatic lexical and syntactic analysis are in operational status; they will now be subjected to comprehensive testing through analysis of English, German, Russian and Chinese texts. Completion of all programs involved in translation by syntactic transfer are scheduled for completion in July 1964. Programs for translation by semantic transfer will be finished by July 1965.

## 5 PROGRESS IN THE QUARTER

### 5.1 Systems

Work done during the quarter by the Systems group centered around making final preparations for installation of the IBM 7040 at the Center. The computer became operational toward the end of the quarter and intensive data processing was begun.

#### 5.1.1 Programming

Program conversion for use on the 7040 was essentially completed. All operational 7090 programs except Grammar Display have been tested and made operational on the 7040. Testing of Grammar Display is now underway.

Most of the new routines required for the Control Program have been completed. All card-to-tape, on-line, and tape-to-print routines are now on the CP tape. Each job may be set up as a single deck of control cards or several jobs may be batched if desired. Some improvements are still required in the tape-to-print routine to simplify the task of the computer operator.

The Monolingual Recognition programs, now available for processing linguistic data on the 7040, were modified during the quarter to accommodate more complex data. The

Table 5-1. Current Status of System Programs

PROGRAM	CURRENT STATUS	NO. OF SEGMENTS	NO. COMPUTER LOCATIONS REQUIRED		LINES OF CODING NOT REQUIRING SPACE
			PROG. INSTRUCT.	CONSTANTS MESSAGES	
Control	5	11	6400	3800	1350
Concordance	4	2	1000	100	100
General Sort	4	1	700	50	200
Request Maint.	6	2	1400	300	200
Corpus Maint.	6	1	1800	200	100
<u>Grammar Maint.</u>					
Rule Rev.	4	6	9900	800	300
Prob. Rev.	4	3	4250	300	100
Input Sel.	4	3	2600	300	500
Output Sel.	1				
Display	4	3	4500	450	550
<u>Transfer Maint.</u>					
Mono. Rev.	1				
Inter. Rev.	*3	3	2600	200	300
Mono. Input Sel.	4	3	2600	300	500
Inter. Sel.	*3	2	500	100	100
Mono. Output Sel.	1				
Mono. Dis.	1				
Inter. Dis.	1				

\*Not converted to 7040

Table 5-1 (continued)

PROGRAM	CURRENT STATUS	NO. OF SEGMENTS	NO. COMPUTER LOCATIONS REQUIRED		LINES OF CODING NOT REQUIRING SPACE
			PROG. INSTRUCT.	CONSTANTS MESSAGES	
<u>Monolingual Recog.</u>					
Lex. Anal. & Choice) Syn. Anal. & Choice)	4	1	2900	350	250
Sem. Anal. & Choice	1				
Lex. Anal. Display) Syn. Anal. Display)	4	3	2000	400	100
Sem. Anal. Display	1				
<u>Interlingual Recog.</u>					
Lex. Analysis	1				
Syn. Analysis	1				
Sem. Analysis	1				
Lex. Display	1				
Syn. Display	1				
Sem. Display	1				
<u>Transfer</u>	1				
<u>Interlingual Prod.</u>					
Lex. Synthesis	1				
Syn. Synthesis	1				
Sem. Synthesis	1				
<u>Monolingual Prod.</u>					
Lex. Choice & Synthesis	1				
Syn. Choice & Synthesis	1				
Sem. Choice & Synthesis	1				
Output Corpus Display	1				

present programs will handle one or two sentences. However, when an attempt was made to process paragraph-size samples of meteorological text, lexical and syntactic complexity of the text precipitated memory overflow and prevented satisfactory processing. Subroutines have been designed to accommodate the memory overflow conditions.

A concordance program was designed and coded during the quarter. The program isolates and sorts each word of a pre-selected text, within its context, and provides a display of all such contexts in the text.

#### 5.1.2 Current Status of Programs

The current status of System programs is given in Table 5-1. The number of lines of coding listed is that required in the machine language of the IBM 7040. The numbers in the first column indicate program status as follows:

1: Planned--Programming specifications have been completed to the level of description of general program objectives, including the input-output and internal data units and the logical operations of the algorithm to be performed.

2: Flowcharted--Programming specifications have been put into the form of data formats and flowcharts having sufficient information to be used as the complete basis for coding.

3: Coded--Programming specifications have been described in the programming language of the computer to be used in implementing the algorithm.

4: Operational--The program has been tested and found to work with small data samples specially designed to verify coding.

5: Completed--The program has been tested and found to work with comprehensive, so-called "real" data, and all inadequacies found in testing have been eliminated.

6: Documented--Documents which describe the algorithm, the program structure which implements it, and conventions for using the program have been prepared for publication.

### 5.1.3 Operations

The Operations section continued to be occupied primarily with providing support for the programmers' efforts in converting the programs for use on the 7040.

The computer was officially turned over to LRC on 15 January. Production processing was begun on 27 January. Jobs were started in Request Maintenance and Rule Revision.

### 5.2 Descriptive Linguistics

The Descriptive Linguistics group continued to compile and revise linguistic data for processing when the



Center's computer became available for use. Tree diagrams were reappraised for conversion to binary form. A study designed to investigate grammar convergence was made during the quarter which yielded guidelines that are being followed in developing grammars for the several languages.

#### 5.2.1 English

Some first statistics on grammar convergence were obtained by processing a special test corpus compiled from an article on meteorology [8]. The text consisted of some 3,000 running words, from which 15 sentences were selected at random for linguistic analysis. Three sub-grammars were provided by dividing the source sentences into three groups and assigning a different range of identification numbers to the rules in each sub-grammar. This permitted one processing run to be evaluated as three separate runs, each with a larger grammar. Rules derived for the first sub-grammar,  $G_1$ , totaled 206; for  $G_2$  an additional 117 rules; and for  $G_3$  an additional 57 rules.

These sub-grammars were used in the computer to process the test corpus, excluding the paragraphs which contained the 15 source sentences.  $G_1$  successfully processed approximately 28% of the text;  $G_1$  combined with  $G_2$  approximately 39%; and all three sub-grammars together processed approximately 44% of the text.

On the basis of these results, it was decided that all grammars will be developed along the following general lines. Ten source sentences will be selected at random from a corpus. A constituent analysis of the sentences will be made and the structures will be classified. After the rules have been encoded and compiled, the grammar will first be checked against a special research corpus consisting of the source sentences. Then the research corpus will be enlarged by adding another set of randomly selected sentences. After automatic analysis is performed, new rules will be added only as necessary to achieve complete analysis. This procedure will be repeated until a generally effective grammar has been built up.

Ten sentences were selected by the English section from Corpus 5 for constituent analysis. The rules derived from these sentences were encoded and listed on a Request Maintenance display. Revisions are being made and the data will be ready for analysis testing early in the next quarter.

Among several approaches to constituent analysis investigated during the quarter, a regressive description of the noun phrase was found to be particularly economical. However, subsequent investigation demonstrated the point often raised in research at the Center: in a hierarchical system of grammars, an optimal collection of rules in an isolated sub-grammar will not necessarily be optimal within a system of several orders of grammar.

Revision was begun of the Request Maintenance display of data from Webster's New Collegiate Dictionary. Membership of all classes are being checked for correct assignment, classes are being partitioned to add new grammatical features, and punch errors are being corrected.

A major effort during the quarter consisted of reappraising tree diagrams and syntactic rules for conversion to binary form. Documentation of the new grammar was begun, and will be used as a model for documentation of grammars for the other languages.

#### 5.2.2 German

The German tree diagrams were revised and syntactic rules were converted to binary form. Initial documentation was begun of the new grammar, patterned on the documentation being prepared for the English grammar.

Ten sentences were selected from Corpus 5 and a constituent analysis was made. The encoded syntactic data was submitted for compilation and display. When the display becomes available, the data will be revised and prepared for analysis testing.

#### 5.2.3 Russian

Research on Russian proceeded along the lines of work with English and German. A pilot study was made of converting Russian tree diagrams and rules to binary form.

Final revision of verb classification format was completed, a task on which effort had been temporarily suspended.

In the next quarter ten sentences will be selected for analysis and processing, along the lines of the grammar development techniques established in the English section.

#### 5.2.4 Chinese

Revision was completed of RG2 requests prepared earlier. A card listing of these requests was proofed and corrected. These data are now ready for compilation and testing.

Conversion to binary form of the tree diagrams and rules was begun. Ten sentences were selected as source sentences for grammar development, and the constituent analysis was completed.

A list was compiled of articles, classifiers and other affixes which will be used as attributes defining formational classes. These data will be used in preparing a concordance display of the Chinese corpus during the next quarter.

#### 5.3. Theoretical Linguistics

Since the general theory of translation underlying the Language Translation System is considered to be completed,

theoretical work is now concentrated upon classification theory and techniques not directly connected with translation research. Future studies of translation processes involving heuristic choices may, however, utilize the results of this investigation.

## CONCLUSIONS

The IBM 7040 at the Center became operational toward the end of the quarter and intensive data processing was begun. Most of the routines needed to adapt the Control Program to console procedures of the 7040 were completed and incorporated onto the CP tape. Subroutines have been designed which will handle overflow conditions that develop when Monolingual Recognition programs are used to analyze large samples of complex data.

The Descriptive Linguistics group has essentially completed a conversion of tree diagrams and syntactic rules to binary form. A grammar convergence study was made during the quarter, from which a procedure was derived for developing optimal grammars for the several languages.

## PLANNING FOR THE NEXT QUARTER

The Systems group will continue testing the converted programs on the IBM 7040 while the programs are being used to process data submitted by the linguists. Programming development on remaining System programs will be resumed.

The Descriptive Linguistic group will be concerned primarily with development of grammars using the techniques which resulted from the pilot study of grammar convergence. Conversion of syntactic rules to binary form will be completed and documentation of the new grammars will be drafted.

## REFERENCES

1. Final Report on Machine Translation Study, Report No. 16. Austin: Linguistics Research Center, June 1963.
2. Symposium on The Current Status of Research, LRC 63-SR1. Austin: Linguistics Research Center, October 1963.
3. N. Chomsky, Syntactic Structures. The Hague: Mouton & Co., 1957.
4. E. Bach, Introduction to Transformational Grammars. New York: Holt, Rinehart and Winston, (in preparation).
5. E. Ruechardt, Sichtbares und Unsichtbares Licht. Berlin: Springer-Verlag, 1952.  
E. Ruechardt, Light, Visible and Invisible. Ann Arbor: University of Michigan Press, 1958.
6. B. Shchetinin, "Ekonomicheskaja Pomoshch' Socialisticheskikh Stran Molodym Nacional'nym Gosudarstvam," Voprosy Ekonomiki, 1960, vol. III, no. 6, pp. 60-70.  
B. Svetlichnyi, "Sovetskoe Gradostroitel'stvo na Sovremennom Etape," Voprosy Ekonomiki, 1960, vol. III, no. 8, pp. 48-60.  
D. Allakhverian, "O Funkcijakj Finansov pri Socializme," Voprosy Ekonomiki, 1960, vol. III, no. 11, pp. 45-55.
7. Yam Chung Ming, [Talks on the Rudiments of Language Teaching]. Hong Kong; 1961.
8. H. Riehl, "On the Origin and Possible Modification of Hurricanes," Science, 1963, vol. 141, No. 3585.



PERSONNEL\*

Administration:

Dr. W. P. Lehmann, Director (82)

Mr. E. D. Pendergraft, Associate Director (236)

Office:

Mrs. Z. Koprivnik, Secretary

Mrs. L. Childress

Mrs. L. Schwab

Descriptive Linguistics:

Dr. L. W. Tosh, Chief (236)

Dr. M. L. Joynes (176)

Chinese:

Mrs. M. Gray

English:

Mrs. M. S. Prince

Mr. K. W. Ryan

German:

Mr. T. Baker

Mrs. K. S. Heilbrunn

Applications:

Dr. A. G. Dale, Chief

Mr. D. W. Allford (176)

Additional Personnel Resumés

Allred G. Dale

Degrees:

B.A., 1951, Oxford University

Ph.D., 1961, University of Texas

Research and teaching:

Lecturer and Assistant Professor of Business Statistics,  
University of Texas, 1953-1963.

Research in computer simulation of decision processes,  
economic model construction, and automatic classification.

Publications:

"Management Decision Tester," Computers and Automation  
(October 1962).

"Developing the Small Business Executive via Simulation,"  
California Management Review (Winter, 1963).

Business Gaming (Bureau of Business Research, University of  
Texas, in press).

Mary Lu Joynes

Degrees:

B.A., 1952, University of Texas

Russian:

Mr. R. B. Bojar (88)

Mr. H. H. Van Olphen

Systems:

Mr. R. W. Jonas, Chief (280)

Operations:

Mr. O. H. Olson, Jr.

Mrs. M. L. Burkland (472)

Mr. D. E. Flatt

Programming:

Mrs. B. C. Foster (472)

Mr. T. W. Hill (88)

Mrs. O. D. Janca (472)

Mr. S. A. McGall (472)

Theoretical Linguistics:

Mr. W. B. Estes, Chief (176)

Mathematics:

Miss J. M. Brady (176)

Mr. D. A. Senechalle (176)

Logic:

Mr. J. Bunnag

Mr. J. Dauwalder

Applications:

Dr. A. G. Dale, Chief

Mr. D. W. Allford (176)

Additional Personnel Resumes

Alfred G. Dale

Degrees:

B.A., 1951, Oxford University

Ph.D., 1961, University of Texas

Research and teaching:

Lecturer and Assistant Professor of Business Statistics,  
University of Texas, 1953-1963.

Research in computer simulation of decision processes,  
economic model construction, and automatic classification.

Publications:

"Management Decision Tester," Computers and Automation  
(October 1962).

"Developing the Small Business Executive via Simulation,"  
California Management Review (Winter, 1963).

Business Gaming (Bureau of Business Research, University of  
Texas, in press).

Mary Lu Joynes

Degrees:

B.A., 1952, University of Texas

M.A., 1955, University of Texas

Ph.D., 1958, University of Texas

Research and teaching:

Instructor and Assistant Professor of English, University  
of Wisconsin, 1957-1961.

Linguistic consultant, English Language Exploratory Committee  
(Tokyo), 1958.

Research linguist, Center for Applied Linguistics, 1962.

\* The number of manhours worked in the quarter under  
contract No. DA 36-039 AMC-02162 (E) are indicated  
in parentheses.

## DISTRIBUTION LIST

OASD (R&E), Rm. 3E1065 ATTN: Technical Library The Pentagon Washington 25, D. C.	1
Chief of Research and Development OCS, Department of the Army Washington 25, D. C.	1
Commanding General U.S. Army Materiel Command ATTN: R&D Directorate Washington 25, D. C.	1
Commanding General U.S. Army Electronics Command ATTN: AMSEL-AD Fort Monmouth, New Jersey	3
Commander, Defense Documentation Center ATTN: TISIA Cameron Station, Bldg. 5 Alexandria, Virginia 22314	20
Commanding General USA Combat Developments Command ATTN: CDCMR-E Fort Belvoir, Virginia	1
Commanding General U.S. Army Materiel Command ATTN: AMCRD-RP-PE, Attn Mr. A. Maklin Bldg T-17 Gravelly Point, Virginia	2
Commanding Officer USA Communication and Electronics Combat Development Agency Fort Huachuca, Arizona	1
Commanding General U.S. Army Electronics Research and Development Activity ATTN: Technical Library Fort Huachuca, Arizona	1
Chief, U.S. Army Security Agency Arlington Hall Station Arlington 12, Virginia	2

Deputy President U.S. Army Security Agency Board Arlington Hall Station Arlington 12, Virginia	1
Director, U.S. Naval Research Laboratory ATTN: Code 2027 Washington 25, D.C.	1
Commanding Officer and Director U.S. Navy Electronics Laboratory San Diego 52, California	1
Aeronautical Systems Division ATTN: ASAPRL Wright-Patterson Air Force Base, Ohio	1
Air Force Cambridge Research Laboratories ATTN: CRZC L.G. Hanscom Field Bedford, Massachusetts	1
Air Force Cambridge Research Laboratories ATTN: CRXL-R L.G. Hanscom Field Bedford, Massachusetts	1
HQ, Electronic Systems Division ATTN: ESAT L.G. Hanscom Field Bedford, Massachusetts	1
Rome Air Development Center ATTN: RAALD Griffiss Air Force Base, New York	1
AFSC Scientific/Technical Liaison Office U.S. Naval Air Development Center Johnsville, Pennsylvania	1
USAE LRDL Liaison Office Rome Air Development Center ATTN: RAOL Griffiss Air Force Base, New York	1
NASA Representative Scientific and Technical Information Facility P. O. Box 5700 Bethesda, Maryland 20014	1

Defense Intelligence Agency ATTN: DIARD Washington, D.C. 20301	1
Commander U.S. Army Research Office (Durham) Box CM-Duke Station Durham, North Carolina	1
Commanding Officer U.S. Army Electronics Materiel Support Agency ATTN: SELMS-ADJ Fort Monmouth, New Jersey	1
Director, Monmouth Office U.S. Army Combat Developments Command Communications-Electronics Agency Fort Monmouth, New Jersey	1
Corps of Engineers Liaison Office U.S. Army Electronics Research & Development Laboratory Fort Monmouth, New Jersey	1
Marine Corps Liaison Office U.S. Army Electronics Research & Development Laboratory Fort Monmouth, New Jersey	1
AFSC Scientific/Technical Liaison Office U.S. Army Electronics Research & Development Laboratory Fort Monmouth, New Jersey	1
Commandant U.S. Army Air Defense School ATTN: Command & Staff Dept. Fort Bliss, Texas	1
Commanding Officer Engineer Research and Development Laboratories ATTN: Technical Documents Center Fort Belvoir, Virginia	1
Commanding Officer U.S. Army Electronics Research & Development Laboratories ATTN: Logistics Division Fort Monmouth, New Jersey For R. M. Dunn, SELRA/NP-6	3 or more



<p> <b>Commanding Officer</b>  <b>U.S. Army Electronics Research &amp; Development Laboratories</b>  <b>ATTN: Director of Research</b>  <b>Fort Monmouth, New Jersey</b> </p>	1
<p> <b>Commanding Officer</b>  <b>U.S. Army Electronics Research &amp; Development Laboratory</b>  <b>ATTN: Technical Documents Center</b>  <b>Fort Monmouth, New Jersey</b> </p>	1
<p> <b>Commanding Officer</b>  <b>U.S. Army Electronics Research &amp; Development Laboratories</b>  <b>ATTN: SELRA/NP-6</b>  <b>Fort Monmouth, New Jersey</b> </p>	1
<p> <b>Commanding Officer</b>  <b>U.S. Army Electronics Research &amp; Development Laboratory</b>  <b>ATTN: Technical Information Division</b>  <b>Fort Monmouth, New Jersey</b> </p>	3
<p> <b>Commanding Officer</b>  <b>U.S. Army Electronics Research &amp; Development Laboratory</b>  <b>ATTN: J. Benson, SELRA/XC</b>  <b>Fort Monmouth, New Jersey</b> </p>	1
<p> <b>Commanding Officer</b>  <b>U.S. Army Electronics Research &amp; Development Laboratory</b>  <b>ATTN: SELRA/X</b>  <b>Fort Monmouth, New Jersey</b> </p>	1
<p> <b>Ramo-Wooldridge Corporation</b>  <b>Canoga Park, California</b>  <b>ATTN: Information Systems Dept.</b> </p>	1
<p> <b>Carnegie Institute of Technology</b>  <b>Schenley Park</b>  <b>Pittsburgh, Pennsylvania</b>  <b>ATTN: Dr. Allen Newell</b>  <b>Dr. Alan Perlis</b> </p>	1 1
<p> <b>R.C.A. Data Systems Division</b>  <b>4922 Fairmont Avenue</b>  <b>Bethesda 14, Maryland</b>  <b>ATTN: Dr. Jack Minker</b> </p>	1
<p> <b>Commanding Officer</b>  <b>U.S. Army Signal Corps School</b>  <b>Officers Dept.</b>  <b>Fort Monmouth, New Jersey</b>  <b>ATTN: ADPS Committee</b> </p>	1

Commanding Officer U.S. Army Signal Center and School Dept. Specialist Training Fort Monmouth, New Jersey ATTN: Information Processing Branch	1
Commanding Officer U.S. Army Signal Engineering Agency Washington 25, D. C. ATTN: Computer Systems Division	1
Lincoln Laboratories Lexington 72, Massachusetts ATTN: Army Liaison Officer	1
National Bureau of Standards Washington 25, D. C. ATTN: Mrs. Ida Rhodes Mechanical Translation Project	1
ATTN: Mr. Russell Kirsch Data Processing Systems Division	1
National Science Foundation Washington 25, D. C. ATTN: Mr. Richard See Documentation Research Program	1
Director, National Security Agency Fort George G. Meade, Maryland ATTN: C-3141 (Rm 2C087) Librarian	2
Professor E. de Grolier Institute National des Techniques de la Documentation Paris, France	1
International Electric Corporation Box 285 ATTN: J. Harlow Paramus, New Jersey	1