

GEORGETOWN AUTOMATIC TRANSLATION

General Information and Operation
Manual

by

J. A. Moyne

Copyright, 1962
Georgetown University
Machine Translation Research
Washington. D. C.
USA

Foreword

The Georgetown Automatic Translation * (GAT) is a result of group effort and cooperation of linguists and programmers. This publication is intended to serve as a manual of operations and reference for the programming system. The system is often referred to as the "Direct Conversion" system since it started as an undertaking for converting the running Georgetown MT system of the IBM 705 computer into the IBM 709 - 7090 computer system. In the course of "conversion", however, several major and minor programs were added to the system and almost all the existing ones were revised and modified.

All the programs described in this manual are available in Squeeze (SOS Share System) or in absolute binary cards which require no compiler. A system tape is also available for the complete GAT system but the latter will only work with the SOS compiler. All inquiries and requests for program material should be addressed to:

Leon E. Dostert
Director
Georgetown University
Machine Translation Research
1330 New Hampshire Avenue, N.W.,
Washington 6, D. C.
United States of America

* Formerly, General Analysis Technique.

Table of Contents

	<u>Page</u>
Foreword	i
0. Machine Translation	1
0.0 Outline of Operations	1
0.1 Speed and Capacity	2
0.2 The GAT Translator	2
1. The GAT Dictionary	3
1.0 The Split Glossary	3
1.1 The Unsplit Glossary	6
1.2 The Dictionary Program and Forms	9
2. Morphology and Look-Up	11
2.0 Input-Output-Procedures	11
2.1 Loading Instructions	11
2.2 The Program Operations	13
3. The GAT Translator	16
3.0 General Descriptions	16
3.1 Load Instructions	17
3.2 Output Print	19
3.3 Linguistic Function of the GAT Translator	22
3.3.0 Routine 01, Sentence Build-up	22
3.3.1 Routine 02, Exclusion	22
3.3.2 Routine 03, Idiomatic	22
3.3.3 Routine 04, Gap Analysis	22
3.3.4 Routine 05, Interpolation	23
3.3.5 Routine 06, Syntagmatic I	23
3.3.6 Routine 07, Syntagmatic II	23
3.3.7 Routine 08, Subject	24
3.3.8 Routine 09, Syntax	24
3.3.9 Routine 10, Lexical Choice	24
3.3.10 Routine 11, Verb Synthesis	24
3.3.11 Routine 12, Noun Synthesis	24
3.3.12 Routine 13, Rearrangement	24
4. The Utility Programs	28
4.0 1401 Dictionary Creation	28
4.1 709 Text Deblocking	29
4.2 1401 GATIP Program	29
4.3 1401 GATOP Program	30
4.4 709 Print Program	32
4.5 1401 Sentence Structure Print	32
4.6 1401 Print Program	32
4.7 709 Sort Program	33
4.8 709 Dictionary Update	35

0. Machine Translation

0.0 Outline of Operations

The GAT system of Machine Translation is comprised of the following operations from the input or source language to the output or target language in the order listed below:

1. The Russian text is keypunched on IBM cards in continuous form; that is, words are punched in the order of the text using columns 2-72 of each card with a blank space between each word (see Appendix I for punching instructions).
2. The punched cards are put on a magnetic tape by a special 1401 computer program which will give a serial number to each word in the text, according to the order of its occurrence in the text, and prepare records in a format suitable for input for a 709, 7090 or 7094 computer.
3. The text on tape is sorted alphabetically,
4. The text is run in a 7090 computer against the GAT Dictionary. Each text word is looked up, and the English meaning or meanings with codes designating the morphological and functional characteristics of each text word are placed in the storage areas reserved for each word.
5. The text is resorted into the original text order.
6. Each Russian sentence is analyzed as one unit and new codes pertaining to the contextual characteristics of each word, phrase, clause, or sentence are generated by the computer and placed in the areas reserved for this purpose.
7. The English translation of each sentence is extracted and organized for printing.
8. Finally the translation is printed by a 1401 computer program which provides for printing the English translation alone or dividing the page into two columns for printing the Russian text and the English translation side by side. This last program has a further option of printing in FORTRAN characters and also leaves room in the printed output for filling in formulae, diagrams, and tables which it is not possible to keypunch owing to the present limitations of keypunch machines.

The two 1401 programs mentioned above are described in the chapter on The Utility Programs; the other programs contained in the above eight steps are described in various other sections of this manual.

0.1 Speed and Capacity

The translation operations of the GAT system (the GAT Translator) contain approximately 35,000 computer instructions (709/7090) with an average of 2 1/4 cycles per instruction or 4.90 microseconds on the 7090 and 4.50 microseconds on a 7094 per instruction. Thus the maximum time required for the internal translation of one Russian sentence, having up to 65 Russian words and containing all the logically possible variations of ambiguities and complexities, is about 17/100 of a second on a 7090 computer. In practice this time is considerably less since all sentences do not have the maximum capacity of 65 Russian words and no sentence has all the logical possibilities combined in one. On the other hand, the speed of the printed output is greatly reduced through input-output delays and tape handlings.

The speed of the preparation and printing of the text depends on the type and size of the 1401 installation. The speed of the two sortings depends on the size of the text and the number of tape units available for the 7090 computer. The programs are under continuous revision for technical improvements which will increase the input-output and internal speeds,

0.2 The GAT Translator

The Translator performs the principal function of the GAT system by analyzing the Russian sentence and composing its English translation. This task is entirely done by a 709, 7090 or 7094 computer. The Term "system" used in connection with the GAT, however, does not imply a compiler or interpreter system as used in computer terminology.

The Translator is merely composed of thirteen routines with a varying number of sub-routines in each routine, all written in the SCAT Mnemonic codes. The routines are originally compiled through the SOS system. Once a routine is debugged absolute binary cards are created and the routine thus becomes independent of any compiler or interpreter. SOS squeeze decks are, however, kept up-to-date with each binary deck to take optimum advantage of the facilities available in the SOS system for making changes and revisions in a program.

Anyone with the knowledge of the language of 709 computer, for example, can read, revise, and operate all the programs of the GAT without need for the study and comprehension of any other system.

The thirteen routines of the GAT Translator are contained in steps 5 and 6 under Section 0.0 above and are further described on the following pages.

1. The GAT Dictionary

The GAT dictionary is composed of two parts: split glossary, and un-split glossary. The former contains the stems of Russian words without any inflectional suffixes; the latter contains Russian words which are not subject to inflection and certain inflected forms.

1.0 The Split Glossary

The split glossary contains all inflected stems or base forms for nouns, adjectives, pronouns, participles, and verbs in Latin characters transliterated from the Cyrillic.*

Each Russian stem in this glossary occupies 34 machine words having 6 characters or 36 bits each. The 34 machine words make up one record (item). The magnetic tape containing this glossary is, however, blocked so that each record on the tape contains 150 items or 5,100 machine words. The 34 machine words in each item is filled with the Russian word, identification codes, the English meaning or meanings, and reserved areas. If the 34 words are assumed to be numbered from left to right, the following table will show the contents of each machine word. The table further demonstrates the format of the split glossary on tape. The machine words on the tape are not numbered and are recorded in continuous form with the appropriate record gaps. Any machine words not occupied by codes are filled with zeroes except the areas for the English meanings where the un-occupied locations are left blank (Fig. 1).

* A transliteration table is included in Appendix I. Coding procedures for the dictionary, a definition of "stem", and discussions concerning mobile vowels, phonological mutations, homographs, and other problems concerning the making of the dictionary are described in Manual for Coding, Georgetown University, Occasional Papers on Machine Translation, 1961.

Table I

<u>Machine Words</u>	<u>Contents</u>
1-5	Russian word. This starts with machine word 1 beginning in the left location; any words or parts of words left blank are filled with zeroes.
6	Locations 1-3 in this word are occupied by the remaining characters of the Russian word, if any, or by zeroes. Location 4 can be zero or "N"; the latter indicating that there are two identical stems (homographs) in the glossary, one of which is a verb stem and the one with the "N" a nominal or adjectival stem. Locations 5 and 6 in this word are zeroes.
7	Locations 1-5 are zeroes; location 6 can contain one of the following numbers indicating the parts of speech specified in the following list: <ul style="list-style-type: none"> 0 - numeral or punctuation 1 - noun 2 - verb 3 - adjective 4 - adverb 5 - preposition 6 - conjunction 7 - particle
8	Locations 1-3 are zeroes; locations 4-6 contain complementary codes for the parts of speech coded in word 7.*
9	Location 1 is zero. Locations 2-6 contain codes which indicate the inflectional types of nouns, adjectives, or verbs as well as codes for the English synthesis.

* For detailed discussions concerning these codes and others specified in this table see the GAT linguistic papers in the series Occasional Papers on Machine Translation, Georgetown University, Washington, D. C., see also Appendix II.

Machine
Words Contents

- 10 Locations 1 and 2 are zeroes. Location 3 gives the number of machine words completely filled by the Russian stem entry; locations 4 and 5 are zeroes; location 6 gives the number of characters in the last machine word partially filled by the Russian stem entry.*
- 11 Locations 1, 2, 4, and 5 are zeroes. Locations 3 and 6 contain codes concerning the exclusion boundaries and candidates used by the Exclusion, Lexical Choice, and other routines.
- 12 Locations 1-3 contain primary idiom codes. Locations 4-6 contain secondary idiom codes.
- 13 Location 1 contains zero. Locations 2-6 contain case determiner codes for strong government.
- 14 Location 1 contains zero. Locations 2-6 contain case determiner codes for weak government.
- 15 Locations 1 and 2 contain particle codes; location 3 contains codes pertaining to the voice or other characteristics of the Russian verb; location 4 can carry a code indicating a verb of motion; location 5 can contain codes pertaining to the translation of the instrumental case; location 6 can contain codes, used by the lexical choice routine, for items which are semantically related to time.
- 16 Locations 1-5 contain zeroes; location 6 can contain codes pertaining to the composition of the English nouns and adjectives ("synthesis" codes).

* For example, if the Russian stem entry in the glossary is DE1STVITEL6-NOST, the contents of machine word 10 would be 002003 indicating that the Russian stem occupies two machine words (having six characters each) and three additional characters in the third machine word (i.e., the Russian stem has 15 characters). This information is needed by the routines in the GAT Translator.

Machine
Words

Contents

- 17 Locations 1-5 contain zeroes; location 6 can contain codes pertaining to the English verbs which require the insertion of "be" (e.g., "be necessary") or items which require the use of "an" as the indefinite article.
- 18 Location 1 contains zero; locations 2-6 contain lexical choice codes and numbers.
- 19 Locations 1-5 contain zeroes; location 6 contains codes pertaining to insertion of the English article.
- 20 Locations 1-5 contains zeroes; location 6 contains codes pertaining to the rearrangement of the Russian words and phrases into the English order.
- 21-26 Area for the first English meaning. The last character of the English word is located in the last location of machine word No. 26. Any machine words or parts of words not occupied by the English meaning are left blank (not zeroes).
- 27 Blank word (used by later routines for "English synthesis" insertions) .
- 28-29 "Overflow" area for the first English meaning (e.g., in Fig. 1 the Russian stem VEDA is translated "be in charge"; "be" is placed in machine words 21-26 and "in charge" is placed in machine words 28 and 29).
- 30-32 Area for the second English meaning, if any.
- 33-34 Record mark and gap.

1.1 The Unsplit Glossary

The unsplit glossary contains the items which are not generally inflected as well as certain inflected items.* These include certain classes of numerals and pronouns, adverbs, prepositions, conjunctions, particles; abbreviations, digital numbers, chemical formulae, non-Russian words, and exceptional cases of declensional forms.

* See Manual for Coding, Georgetown University, Occasional Papers on Machine Translation, 1961.

Each item (or entry) in the unsplit glossary occupies 44 machine words. Each record on the unsplit glossary tape consists of 68 items or 68 X 44 = 2,992 machine words.

Machine words 1-32 in the unsplit item are occupied by the same codes and characters as in the split item given in Table I under 1.0. The possible contents of machine words 33-44 are given in Table II below (see also Fig. 2).

Table II

<u>Machine</u>	
<u>Words</u>	<u>Contents</u>
21-28	Area for the first English meaning.
29-34	Area for the second English meaning.
33	Codes pertaining to the singular cases of the item.
34	Codes pertaining to the plural cases of the item.
35	Locations 1-5 contain zeroes; location 6 contains codes pertaining to the number of the Russian item (i.e., singular, plural, or both) .
36	Locations 1-5 contain zeroes; location 6 contains codes pertaining to the gender of the item.
37	Locations 1-5 contain zeroes; location 6 contains codes pertaining to the form of the adjective or participle (long or short forms).
38	Locations 1-5 contain zeroes; location 6 contains codes pertaining to the animateness of the item.
39	Locations 1-5 contain zeroes; location 6 contains codes pertaining to the voice of the item.
40	Locations 1-5 contain zeroes; location 6 contains codes pertaining to the mood of the item.
41	Locations 1-5 contain zeroes; location 6 contains codes pertaining to the tense of the item.

Machine
Words

Contents

42	Locations 1-5 contain zeroes; location 6 contains codes pertaining to the person of the item.
43-44	Record mark and gap.

Any locations in the above table not occupied by codes are filled with zeroes, except in the areas for the English meanings where the unoccupied locations are left blank.

1.2 The Dictionary Programs and Forms

The programs for the creation and up-dating of the GAT Dictionary are described under The Utility Programs. The Split and Unsplit glossaries are available on separate tapes. All dictionary and system tapes are in high density. Figures 1 and 2 show the format of several Russian words (items) in both glossaries. A complete printed listing of these glossaries is also available. The glossaries are originally coded by hand on special forms. The coding instructions are contained in the Manual for Coding.

2. Morphology and Look-Up

This program is available both as an SOS squeeze deck or an absolute binary deck.

2.0 Input - Output Procedures

The input to this program is the Russian text after it has run through the 1401 preparation program (GATIP) and has been sorted alphabetically. Each item in the text occupies seven machine words. The first word is taken by the serial number and the remaining six are filled with the Russian word beginning from the left most location of machine word two and filling any unoccupied words or parts of words with zeroes. The GATIP program (full descriptions under The Utility Programs) has several options for blocking; normally the output from GATIP is set for 10 items per record or $10 \times 7 = 70$ machine words per record. The output from GATIP is, however, used as input to the alphabetic sort. The output from the sort is blocked for 100 items per record or $100 \times 7 = 700$ machine words per record. This latter output is the input to the morphology program.

The output from morphology consists of records of 45 machine words per item. The contents of these words are the same characters and codes defined in Tables I and II plus one word at the beginning containing the serial number of the text (Fig. 3). This output is in alphabetic order and has to be sorted according to the serial number before it can be used as input to the GAT Translator. Figure 3 is a sample page of the morphology output after it has been sorted.

The morphology program produces two output tapes: one contains the morphology output as outlined above and the other contains a list of the Russian words in the text which were not found in the glossaries. This list can be used to update the glossaries and enrich their contents.

2.1 Loading Instructions

2.1.0 When using the absolute binary deck:

1. Place the absolute deck (it has its own loader program) in the card reader.
2. Mount split glossary on Tape Unit 6, Channel A; mount unsplit glossary on Tape Unit 5, Channel A.

3. Mount the text on Tape Unit 6, Channel B.
4. Mount output tapes on Unit 5, Channel B, and Unit 4, Channel A. Morphology output will be on B-5, and error list (or list of words not found in the glossaries) will be on A-4 (Fig. 4).
5. Clear the computer, ready the card reader, ready the on-line printer, and press LOAD CARD.
6. In the course of processing follow instructions printed by the on-line printer.

2.1.1 When using the SOS squoze deck:

1. Mount SOS system tape on Unit 1, Channel A.
2. Place the squoze deck in the card reader and ready the machine.
3. Mount work tapes on Unit 2, Channel A, and Units 1 and 2, Channel B.
4. Mount output tapes on Unit 5, Channel B, and Unit 4, Channel A.
5. Mount the text on Unit 6, Channel B.
6. Mount split glossary on Unit 6, Channel A, and unsplit glossary on Unit 5, Channel A.
7. Put switch 1 (sign switch) down.
8. Clear the computer and press LOAD TAPE.

2.1.2 When using the SOS squoze on a tape.

Do same as in 2.1.1 omitting steps 2 and 7 and mounting the tape containing the morphology program (squoze) on Unit 3, Channel A.

Figure 4 shows a schematic representation of the input-output units for the morphology and look-up program.

2.2 The Program Operations

The morphology and look-up program performs the basic task of looking up the text words in the glossaries and copying out their English equivalents and pertinent codes as well as generating new codes discernible by

the nature of the suffixes and other morphological factors in the text word, This program has about 4,000 computer instructions and can be run on a 709, 7090, or 7094 computer. The input, output, and internal processing operations are "buffered" and overlapped such that maximum speed is maintained in the running of the program.

The split glossary is read into three areas, A1, A2, and A3, in the memory - - each area can take 150 items or 5,100 machine words. The unsplit glossary is read into two areas, B1 and B2, in the memory - - each area can take 68 items or 2,992 machine words. The input text is read into two areas, C1 and C2, in the memory - - each area can take 100 items or 700 machine words. There is also a work area (starting with word TWORK) and an output area (starting with word OUTPT) of 45 machine words each. While the processed text is being written on tape from the output area new portions of the text and the glossaries are read into the alternating areas and processing is carried out in the other area. As the text and the glossaries are in alphabetic order (709 sequence) each batch of the text is processed against a certain portion of the glossaries and there is no need to rewind and reread the glossary tapes. However, when there is a tape redundancy check or certain portions of the split glossary may have advanced too far in the process of cutting the endings (see below), the split glossary is backspaced appropriately.

The morphology program checks first for reflexive verbs and unloads appropriate codes when these are found (see the list of codes in Appendix II). Following this operation a search is made in the unsplit glossary for the text word. If found, codes are unloaded; if not found, a search is made in the split glossary for the stem. A stem may have zero, one, two, or three characters as suffix or ending. When the stem has been found, the appropriate table of endings contained in the morphology program, is looked up for the codes pertaining to the suffix.

When the above operations fail to find a text word in the dictionary the Russian word is placed in the English meaning area of the output item and digit 8 is unloaded in machine word 8 (Parcue) to indicate this (Fig. 3). Thus, when a Russian word is not found in the dictionary, it is printed in Russian (transliterated) in the translation text. The unfound word is also written on the tape (A-4) provided for the "error list".

Finally the output tape on Unit 5, Channel B, is run through the 709 sort based on the serial number contained in the six characters of the first machine word of the output item. The text is thus put back into the original sequence before any further operations are effected. The control cards for the sort and descriptions are included in Section 4, The Utility Programs.

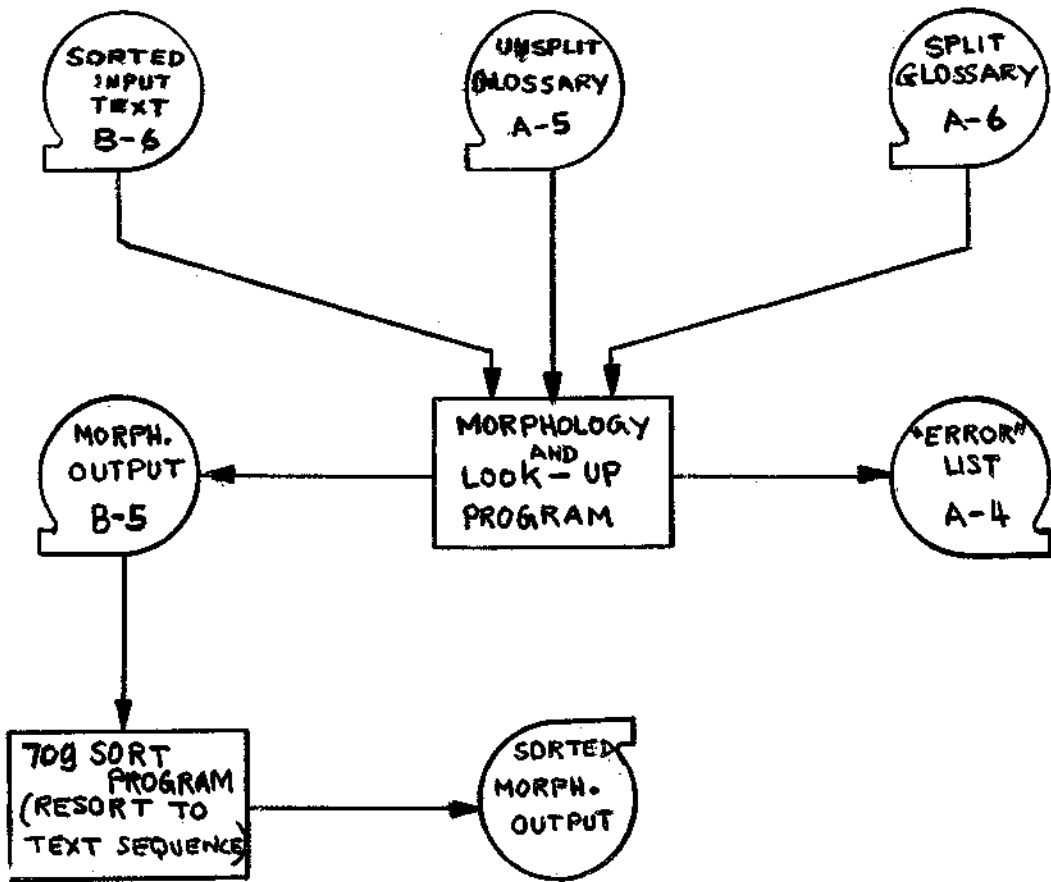


Figure 4

Morphology and Look-Up Program
 Text preparation for input to the GAT Translator

3. The GAT Translator

3.0 General Descriptions

This program contains thirteen routines and a number of subroutines. Absolute binary decks and SOS squeeze decks are available for each one of the routines. Each routine can be run independently; output can be produced and printed for linguistic study, corrections, modifications, and program debugging. The output from each routine becomes input to the next routine. Table III gives a list of the input-output units for the routines in this program.

For production purposes and in the interest of speed the thirteen routines of this program can be arranged into three computer passes. Figure 5 is a schematic representation of the arrangements for these passes and the input-output tape units for each pass.

Table III

<u>Order of Run</u>	<u>Name of Routine</u>	<u>Input tape Unit</u>	<u>Output tape Unit</u>	<u>TRA</u>
1	Sentence Build-up	A-4	B-8	
2	Exclusion	B-8	A-8	
3	Idiomatic	A-8	B-8	
4	Gap Analysis	B-8	A-8	
5	Interpolation	A-8	B-8 & B-6	
6	Syntagmatic I	A-8	B-8 & B-6	
7	Syntagmatic II	A-8	B-8 & B-6	
8	Subject	A-8	B-8 & B-6	
9	Syntax	A-8	B-8 & B-6	
10	Lexical Choice	B-8	A-8 & B-6	
11	Verb Synthesis	A-8	B-8 & B-6	
12	Noun Synthesis	A-8	B-8 & B-6	
13	Rearrangement	A-8	B-8 & B-6 and A-4	

The TRA Column in Table III can contain a transfer address for the following purpose:

In the course of translating a new text it is possible (although rarely) that a new situation or conditions may cause the particular routine to hang-up or go into an endless loop. If it is not desirable at that moment to stop the processing to find the cause of difficulty, the translation operation can be continued by manual transfer to the address in the TRA column. This will generally cause the routine to write out the record which has caused the difficulty and start processing the subsequent records. The transfer addresses have not, however, been included in this manual and column TRA is left blank because each time a routine deck is updated or modified these addresses would change. Anyone intending to use the system should ask for the transfer address for each routine at the time of obtaining the program decks and insert the transfer address in the TRA column by hand for reference.

The following example will serve to illustrate the use of the transfer address. If the transfer address for routine No. 06 (Syntagmatic I) is 33761 and the routine hangs-up in the middle of a text processing, do the following to continue processing.

1. Put the computer in the MANUAL mode.
2. Put key No. 7 down and enter the transfer address 33761 in the address of the keys.
3. Press ENTER INSTRUCTION.
4. Reset the keys and put the computer in the AUTOMATIC mode.
5. Press START.

3.1 Load Instructions

A. If the absolute binary decks are used, the following procedure should be observed for running each routine (or pass when the routines are combined into the three passes).

1. Place the absolute deck in the card reader and make the machine ready (each deck has its own loader program in front of it).
2. Mount the input-output tapes as defined in Table III or Figure 5.
3. Clear the computer and press LOAD CARD.

When the processing of the input tape is completed a message is printed to that effect by the on-line printer.

The output tape for routines Number 5-13 does not rewind automatically to permit processing of more than one tape of input and writing on the same output tape. Therefore, when the processing in one routine is completed, and before going to the next routine, the operator should manually write an end-of-file on the output tape and rewind it.

B. When squeeze decks are used for the continuous run of all the thirteen routines, it is more efficient to put the decks on tape and create a system tape. The squeeze decks on the tape should follow each other without a file gap between the routines. The following format is used on the system tapes which are available for distribution. Thus the system tape contains one file only.

```

Routine 01      JOB      01
                LG
                (preface)
                MOD
                (alters, if any)
                ENDMOD
                (blank)
                (squeeze deck)
                (blank)
                GO
                PAUSE
                (blank)
                JOB      02
                .
                .
                .
                (the same format repeated)
Routine 02      .
                .
                .
Routine 13      (blank)
                (end-of-file)
```

To load and operate this system tape the following steps should be taken:

1. Mount SOS tape on Unit 1, Channel A.
2. Mount GAT Translator system tape on Unit 3, Channel A.
3. Mount work tapes on Unit 2, Channel A, and Units 1 and 2, Channel B.
4. Mount the input-output tapes as defined in Table III or Figure 5.
5. Clear the computer and press LOAD TAPE.

After processing each routine the computer will stop. To go to the next routine the following steps should be taken:

1. Rewind tape Unit 1, Channel A.
2. Change the input-output tapes, when necessary.
3. Clear the computer and press LOAD TAPE.
4. If the computer stops on 1412, press START.

End-of-file and end-of-job messages are printed by the on-line printer. If switch 3 is put down alterations in the routines, if any, and irregularity messages are also printed by the on-line printer.

C. When morphology and look-up program is included in the GAT system tape, the tape has two files: file one contains the morphology and look-up program and file two contains the thirteen routines of the GAT Translator.

3.2 Output Print

When it is desirable to print the output of any one of the thirteen routines of the GAT Translator it is necessary to run the output through the 709 Print Program first and print the output from this program by a special 1401 program. The purpose of the 709 Program and the 1401 Sentence Structure Print * is to give names to each machine word. Figures 6 and 7 are sample pages from these outputs. These outputs are designed for easy linguistic and programming checks and corrections. Figure 6 is a sample page of the output from Routine 05 (Interpolation and the four routines before that) of the GAT Translator; Figure 7 is a sample page

* For details of these programs refer to the section on The Utility Programs.

from the output of the last routine (Rearrangement). By comparing these two routines it will be noted that some of the codes have been altered as a result of the operations that have followed Routine 05. When the object of running the routines is to produce a final translation, these intermediate printings are omitted and the thirteen routines are run successively.

The first routine of the GAT Translator (Sentence Build-up) expands the number of machine words from 45 (produced by morphology output; Fig.3) to 55 (Fig. 6) in each item, and this number is kept constant for each Russian word (item) through the rest of the routines. The additional machine words contain the codes generated by the routines in establishing nesting, agreement, government, and other syntagmatic and syntactic strings within a sentence in the course of the operations. Appendix II contains a complete list of the 55 words of each item and their contents.

The information written on tape Unit 6, Channel B, by routines 5-13 is concerned with the linguistic errors, failures, and new situations found by the routines in the course of processing a text; it provides a feedback system for the correction and updating of the routines.

The following example will serve to illustrate the use of the print program and the information on tape Unit 6, Channel B. Let us suppose it is desirable to print the output from Routine 08 and study the messages on tape B-6 generated by this and other routines preceding this routine; the following steps should be taken:

1. The output from routine 08 is on tape Units 8 and 6, Channel B. Manually put end-of-file on B-8 and B-6 and rewind the tapes.
2. Place the absolute deck for the 709 Print Program in the card reader and ready the machine.
3. Mount output tape on Unit No. 8, Channel A.
4. Clear the computer and press LOAD CARD.
5. After the information on tape Unit 8, Channel B, is processed and copied on tape Unit 8, Channel A, the computer will print a message by the on-line printer to that effect and will halt. Press START to copy the information on B-6 also on A-8.
6. Take tape Unit 8, Channel A, and print it with the special 1401 Print Program.

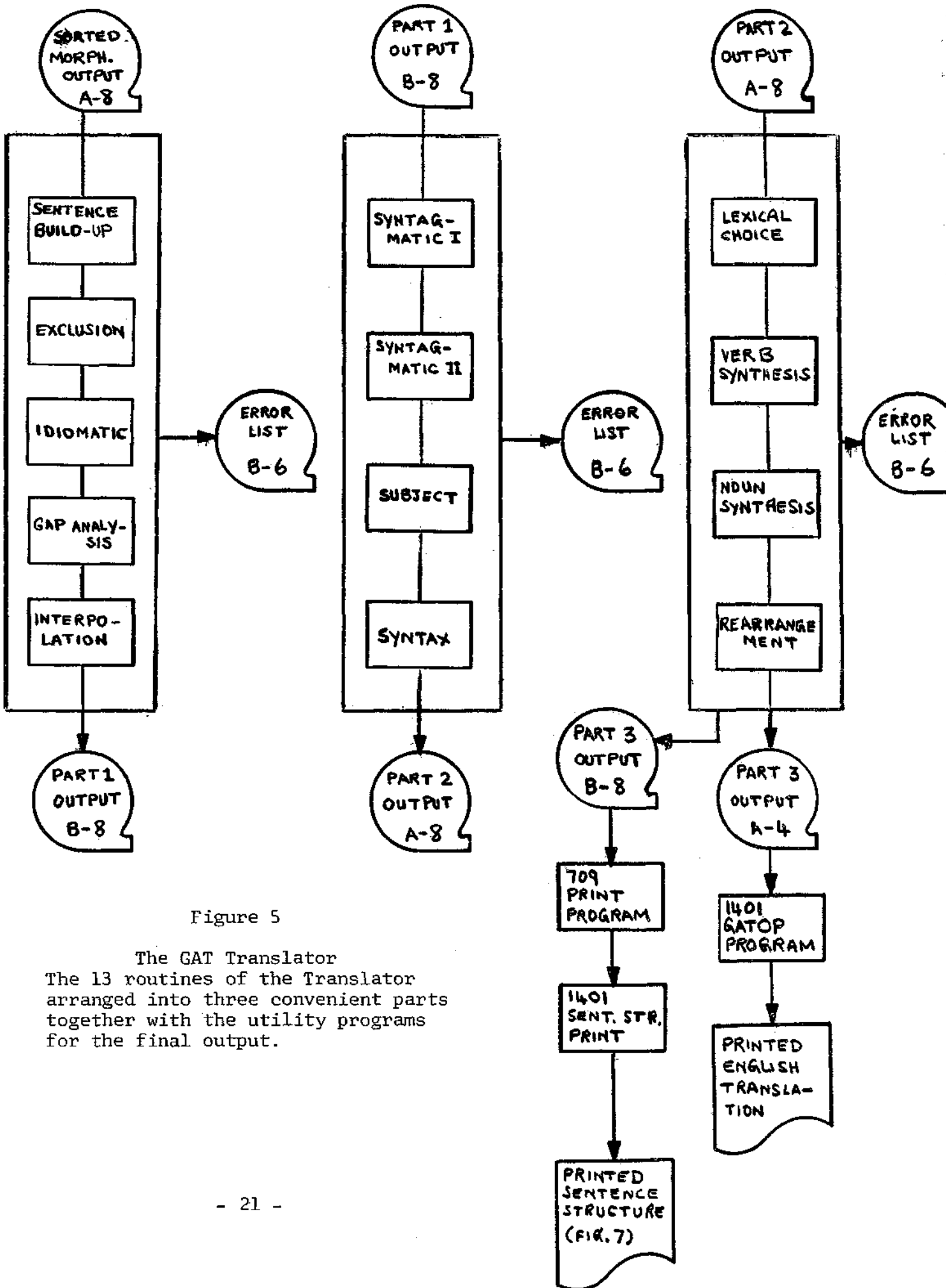


Figure 5

The GAT Translator

The 13 routines of the Translator arranged into three convenient parts together with the utility programs for the final output.

3.3 Linguistic Function of the GAT Translator

The logical and linguistic function of each routine of the GAT Translator is given below in brief outlines for a general understanding of the system. Details of these routines have been given elsewhere,*

3.3.0 Routine 0.1, Sentence Build-up.

This routine expands the 45 words of each item from the morphology output to 55 words, reserving locations for additional codes to be generated by the subsequent routines. It organizes the sentence format which is acceptable by the subsequent routines and establishes sentence boundaries. If any sentence is longer than the maximum allowable length of 65 Russian words, the routine makes the necessary cut or cuts creating multiple sentences from the one long sentence.

3.3.1 Routine 02, Exclusion.

The main function of this routine is to delimit strings within a sentence or a text which are not necessarily subject to linguistic analysis by the later routines. Such "excluded" segments are formulae, diagrams, figures, tables of numbers, and other similar contents of a scientific text.

3.3.2 Routine 03, Idiomatic.

This routine contains tables of idioms in the Russian language and their English equivalents. A simple arithmetic operation adds the "idiom numbers" included in the relevant items in the glossaries and when the permissible total is reached excludes the idiom string from further analysis and inserts the English equivalents.

3.3.3 Routine 04, Gap Analysis.

The function of this routine is to determine the contextual function of the Russian words which are not found in the glossaries. It was stated above (under 2.2) that when a word is not found in the Dictionary the program unloads digit 8 in machine word 8 which contains the code for parts of speech. In the Gap Analysis routine this word is studied together with its cooccurrences to determine its function in the sentence. If it is determined that it functions as an adjective, for example, digit 8 is deleted from Machine word 8 and is replaced by digit 3 which is the code for an adjective.

* See the series of the GAT papers published by Georgetown University as Occasional Papers on Machine Translation, listed in Appendix III.

3.3.4 Routine 05, Interpolation.

This routine contains several subroutines that examine punctuation marks in a sentence and determine their significance. A sentence is often "cut" into convenient shorter clauses on the basis of the function of punctuation marks. These clauses can be treated independently by the subsequent routines. The routine also contains subroutines dealing with the function of formulae, numbers, abbreviations and certain other classes in a sentence or clause. Codes are generated and unloaded in predetermined locations and new boundaries for sentences are established whenever applicable.

3.3.5 Routine 06, Syntagmatic I.

This routine contains the following subroutines:

3.3.5.0 Preposition Cleaning: The subroutine examines every preposition that may govern more than one case and, on the basis of the contextual functions of the preposition, determines what specific case any particular preposition may govern.

3.3.5.1 Adjective Case & Number Cleaning: The subroutine examines all adjectives and participles that may have more than one case or are ambiguous in their number and determines a simple case and number for each one.

3.3.5.2 Full-Nesting and Semi-Nesting: These subroutines establish nested structures within a sentence and code these structures accordingly.

3.3.5.3 Noun Case & Number Cleaning: These subroutines examine the nouns which may have more than one case or are ambiguous in number and resolve the ambiguity in each case.

3.3.5.4 Agreement: This subroutine establishes agreement between a nucleus noun and its modifiers and delimits and codes such stretches accordingly.

3.3.6 Routine 07, Syntagmatic II.

This routine contains a number of subroutines which establish government relationship between nouns, verbs, prepositions, adjectives, and other admissible classes, and delimits stretches containing such relationship by a series of computer generated codes.

3.3.7 Routine 08, Subject.

This routine establishes the subject or subjects in a sentence and codes them accordingly.

3.3.8 Routine 09, Syntax.

This routine has a number of subroutines that determine the number of subjects and predicates in a sentence and classify sentences on the basis of this number. This routine contains the grammatical treatment of a sentence as a unit and provides for the insertion of the English copula and other similar insertions in the translation.

3.3.9 Routine 10, Lexical Choice.

This routine is designed to treat the Russian words that may have more than one English equivalent and make a choice of the one which is most appropriate. Inasmuch as the dictionary can carry only two English meanings for any one Russian entry, this routine contains lengthy tables of further equivalents which may be selected to replace those which are in the dictionary.

3.3.10 Routine 11, Verb Synthesis.

The English equivalent for verbs is entered in glossaries in the infinitive form (without the "to"). This routine puts the English verb in the proper form (past, present, etc.) and inserts the auxiliaries whenever necessary.

3.3.11 Routine 12, Noun Synthesis.

This routine changes the basic forms of English nouns and adjectives into the proper forms dictated by the structure of the sentence (singular, plural, comparative, superlative, etc.). The routine also contains subroutines that insert English prepositions (to, on, in, of, etc.) and definite and indefinite articles (the, a, an) whenever necessary.

3.3.12 Routine 13, Rearrangement.

This routine rearranges the Russian words, and other sentence segments, into the English order for final translation. It further arranges the Russian and English words for the final output. Apart from the error or message tape (B-6), this routine produces two output tapes: tape Unit 8, Channel B, contains the sentence structure

as produced, by all other routines (see 3.2 and Figures 6 and 7) and tape Unit 4, Channel A, which contains the final translation and must be run against the 1401 GATOP program to print the final translation output.

The thirteen routines outlined above can be divided into two groups: Analysis and Synthesis.* The Analysis Group contains Routines 01 to 09 and is principally concerned with the analysis of the Russian language with no significant reference to the target or output language.

The Synthesis Group contains Routines 10 to 15 which are essentially concerned with the composition of the English translation. In order to use the system for translating into any language other than English, it is sufficient to write routines for that language to replace Routines 10 to 13 and link them with the Analysis Group.

* The term "synthesis" has a special usage in the GAT terminology signifying the modification of the basic forms of target language lexical units in accordance with the structure of the sentence in which they occur.

01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19
27483-1						1C2	CCCC	CC605	CCCC	CC002	CCCC	CCCC	CCCC	CCCC	CCCC	CCCC	CCCC	CCCC
SER.	RWD1	RWD2	RWD3	RWD4	RWD5	RWD6	PARC	INTP	FARS	MDLG	EXCL	IDIM	SGVT	WGVT	VSMT	NSAR	BLOC	LEXC
ARTL	RERC	EPRP	DIND	IMG1	IMG2	IMG3	IMG4	IMG5	IMG6	ESYN	CF01	CF02	ZPG1	ZPG2	ZPG3	SING	PLUR	NUMB
SEND	FANV	MCCB	TENS	PERS	NMPF	APLC	AGLC	GVLC	GVLW	GPLC	GNLC	GALC	GVCD	HPLC	DCLC	BLANK		
27484VERXATE																		
SER.	RWD1	RWD2	RWD3	RWD4	RWD5	RWD6	PARC	INTP	FARS	MDLG	EXCL	IDIM	SGVT	WGVT	VSMT	NSAR	BLOC	LEXC
ARTL	RERC	EPRP	DIND	IMG1	IMG2	IMG3	IMG4	IMG5	IMG6	ESYN	CF01	CF02	ZPG1	ZPG2	ZPG3	SING	PLUR	NUMB
SEND	FANV	MCCB	TENS	PERS	NMPF	APLC	AGLC	GVLC	GVLW	GPLC	GNLC	GALC	GVCD	HPLC	DCLC	BLANK		
27485KRVYLC																		
SER.	RWD1	RWD2	RWD3	RWD4	RWD5	RWD6	PARC	INTP	FARS	MDLG	EXCL	IDIM	SGVT	WGVT	VSMT	NSAR	BLOC	LEXC
ARTL	RERC	EPRP	DIND	IMG1	IMG2	IMG3	IMG4	IMG5	IMG6	ESYN	CF01	CF02	ZPG1	ZPG2	ZPG3	SING	PLUR	NUMB
SEND	FANV	MCCB	TENS	PERS	NMPF	APLC	AGLC	GVLC	GVLW	GPLC	GNLC	GALC	GVCD	HPLC	DCLC	BLANK		
27486SF																		
SER.	RWD1	RWD2	RWD3	RWD4	RWD5	RWD6	PARC	INTP	FARS	MDLG	EXCL	IDIM	SGVT	WGVT	VSMT	NSAR	BLOC	LEXC
ARTL	RERC	EPRP	DIND	IMG1	IMG2	IMG3	IMG4	IMG5	IMG6	ESYN	CF01	CF02	ZPG1	ZPG2	ZPG3	SING	PLUR	NUMB
SEND	FANV	MCCB	TENS	PERS	NMPF	APLC	AGLC	GVLC	GVLW	GPLC	GNLC	GALC	GVCD	HPLC	DCLC	BLANK		
27487E/																		
SER.	RWD1	RWD2	RWD3	RWD4	RWD5	RWD6	PARC	INTP	FARS	MDLG	EXCL	IDIM	SGVT	WGVT	VSMT	NSAR	BLOC	LEXC
ARTL	RERC	EPRP	DIND	IMG1	IMG2	IMG3	IMG4	IMG5	IMG6	ESYN	CF01	CF02	ZPG1	ZPG2	ZPG3	SING	PLUR	NUMB
SEND	FANV	MCCB	TENS	PERS	NMPF	APLC	AGLC	GVLC	GVLW	GPLC	GNLC	GALC	GVCD	HPLC	DCLC	BLANK		
27488..																		
SER.	RWD1	RWD2	RWD3	RWD4	RWD5	RWD6	PARC	INTP	FARS	MDLG	EXCL	IDIM	SGVT	WGVT	VSMT	NSAR	BLOC	LEXC
ARTL	RERC	EPRP	DIND	IMG1	IMG2	IMG3	IMG4	IMG5	IMG6	ESYN	CF01	CF02	ZPG1	ZPG2	ZPG3	SING	PLUR	NUMB
SEND	FANV	MCCB	TENS	PERS	NMPF	APLC	AGLC	GVLC	GVLW	GPLC	GNLC	GALC	GVCD	HPLC	DCLC	BLANK		
27489																		
SER.	RWD1	RWD2	RWD3	RWD4	RWD5	RWD6	PARC	INTP	FARS	MDLG	EXCL	IDIM	SGVT	WGVT	VSMT	NSAR	BLOC	LEXC
ARTL	RERC	EPRP	DIND	IMG1	IMG2	IMG3	IMG4	IMG5	IMG6	ESYN	CF01	CF02	ZPG1	ZPG2	ZPG3	SING	PLUR	NUMB
SEND	FANV	MCCB	TENS	PERS	NMPF	APLC	AGLC	GVLC	GVLW	GPLC	GNLC	GALC	GVCD	HPLC	DCLC	BLANK		
27490--																		
SER.	RWD1	RWD2	RWD3	RWD4	RWD5	RWD6	PARC	INTP	FARS	MDLG	EXCL	IDIM	SGVT	WGVT	VSMT	NSAR	BLOC	LEXC
ARTL	RERC	EPRP	DIND	IMG1	IMG2	IMG3	IMG4	IMG5	IMG6	ESYN	CF01	CF02	ZPG1	ZPG2	ZPG3	SING	PLUR	NUMB
SEND	FANV	MCCB	TENS	PERS	NMPF	APLC	AGLC	GVLC	GVLW	GPLC	GNLC	GALC	GVCD	HPLC	DCLC	BLANK		
27491																		
SER.	RWD1	RWD2	RWD3	RWD4	RWD5	RWD6	PARC	INTP	FARS	MDLG	EXCL	IDIM	SGVT	WGVT	VSMT	NSAR	BLOC	LEXC
ARTL	RERC	EPRP	DIND	IMG1	IMG2	IMG3	IMG4	IMG5	IMG6	ESYN	CF01	CF02	ZPG1	ZPG2	ZPG3	SING	PLUR	NUMB
SEND	FANV	MCCB	TENS	PERS	NMPF	APLC	AGLC	GVLC	GVLW	GPLC	GNLC	GALC	GVCD	HPLC	DCLC	BLANK		
27492--																		
SER.	RWD1	RWD2	RWD3	RWD4	RWD5	RWD6	PARC	INTP	FARS	MDLG	EXCL	IDIM	SGVT	WGVT	VSMT	NSAR	BLOC	LEXC
ARTL	RERC	EPRP	DIND	IMG1	IMG2	IMG3	IMG4	IMG5	IMG6	ESYN	CF01	CF02	ZPG1	ZPG2	ZPG3	SING	PLUR	NUMB
SEND	FANV	MCCB	TENS	PERS	NMPF	APLC	AGLC	GVLC	GVLW	GPLC	GNLC	GALC	GVCD	HPLC	DCLC	BLANK		
27493																		
SER.	RWD1	RWD2	RWD3	RWD4	RWD5	RWD6	PARC	INTP	FARS	MDLG	EXCL	IDIM	SGVT	WGVT	VSMT	NSAR	BLOC	LEXC
ARTL	RERC	EPRP	DIND	IMG1	IMG2	IMG3	IMG4	IMG5	IMG6	ESYN	CF01	CF02	ZPG1	ZPG2	ZPG3	SING	PLUR	NUMB
SEND	FANV	MCCB	TENS	PERS	NMPF	APLC	AGLC	GVLC	GVLW	GPLC	GNLC	GALC	GVCD	HPLC	DCLC	BLANK		
27494																		
SER.	RWD1	RWD2	RWD3	RWD4	RWD5	RWD6	PARC	INTP	FARS	MDLG	EXCL	IDIM	SGVT	WGVT	VSMT	NSAR	BLOC	LEXC
ARTL	RERC	EPRP	DIND	IMG1	IMG2	IMG3	IMG4	IMG5	IMG6	ESYN	CF01	CF02	ZPG1	ZPG2	ZPG3	SING	PLUR	NUMB
SEND	FANV	MCCB	TENS	PERS	NMPF	APLC	AGLC	GVLC	GVLW	GPLC	GNLC	GALC	GVCD	HPLC	DCLC	BLANK		
27495																		
SER.	RWD1	RWD2	RWD3	RWD4	RWD5	RWD6	PARC	INTP	FARS	MDLG	EXCL	IDIM	SGVT	WGVT	VSMT	NSAR	BLOC	LEXC
ARTL	RERC	EPRP	DIND	IMG1	IMG2	IMG3	IMG4	IMG5	IMG6	ESYN	CF01	CF02	ZPG1	ZPG2	ZPG3	SING	PLUR	NUMB
SEND	FANV	MCCB	TENS	PERS	NMPF	APLC	AGLC	GVLC	GVLW	GPLC	GNLC	GALC	GVCD	HPLC	DCLC	BLANK		

Figure 6
A sample page of output from Part 1 of the GAT Translator. Codes encircled mark the changes in comparison with Fig. 7.

