## A REPORT LANGUAGE AND

## INTERPRETER FOR DATA ANALYSIS

by
L. D. Lubin

Department of Computer Science
Oniversity of Manitoba
Ginnipeg. Manitoba
May 1974

E Report Language and Interpreter for Data Analysis

A Thesis Presented to<br>The Faculty of Graduate Studies and Research The University of Manitoba Winnipeg, Manitoba

In Partial Fulfilment
of the Requirements for the Degree Master of Science in the Department of Computer Science

## ACKNORIEDGEMENTS

I would like to thank Professor P . H . Dirksen for his advice during the uriting of this thesis. His constructive criticism uas almays a help. I would also like to thank my wife, Ruthann. for her help in proof-reading and typing this thesis and Miss S. Elo of the Oniversity of Toronto Computer Centre who typed the final version.

Thanks are also due to the computing centres of both the University of Manitoba and the University of Toronto for allowing me full access to their facilities.

1. INTRO DUCTION
1.1 Nature and Scope of Thesis
1.2 Organization of Thesis
2. HISTORICAL REVIEW
3. 1 The Problem
2.2 Other Solutions
2.2. 1 SRG-40 SURVEY REPORT LANGUAGE
2.2.2 STATISTICAL REPORI GENERATOR
2.2.3 DATA TEXT
2.2.4 COBOL
2.2.5 API
4. DARE - A DĀTA REPORT LANGUAGE
3.1 Philosophy
5. 2 Language Description
3.2.1 File Definition
3.2.2 Tabulation Section
3.2.2.1 READ KRITE Data Transfer
3.2.2.2 CALCULATE
3.2.2.3 CLASSIFY ORDER PLOT
3.2.2.4 TF
3.2.3 Comment on Examples
6. LANGOAGE IMPLEMENTATION
7. 1 HANDARIN - Manitoba Data Report Interpreter

| 4.1 .1 | Philosophy of Implementation |
| :---: | :---: |
| 4.1 .2 | Gross Logic |
| 4.1.3 | MANDARIN Routines |
|  | 4.1.3.1 CONTROL PROCESSOR |
|  | 4.1.3.2 FILE PROCESSOR |
|  | 4.1.3.3 TABULATE PROCESSOR |
|  | 4.1.3.4 EXECUTION PROCESSOR |

4.2 Interaal organization

| 4.2 .1 | Symbol Table Entries |
| :--- | :--- |
| 4.2 .2 | Instruction Entries |
|  | $4.2 .2 .1 \quad$ READ |
|  | 4.2 .2 .2 MRITE |
|  | $4.2 .2 .3 \quad$ MOVE |
|  | $4.2 .2 .4 \quad$ CALCULATE |
|  | $4.2 .2 .5 \quad$ ISN |
|  | $4.2 .2 .6 \quad$ CLASSIFY/ORDER |
|  | 4.2 .2 .7 |
|  | 4.2 .2 .8 |
|  |  |

5. CONCLOSION

APPENDIX A
APPENDIX B

## INTRODUCTION

### 1.1 Nature_and Scope_of Thesis

The area of computer languages has been and will continue to be a very important area of research in the computer field. We have low-level and high-level languages for almost all types of specialized as well as general purpose applications. We even have languages to help implement other languages. With all of these languages, though, there are still areas where a computing need exists but the solution of a problem is made difficult because of a lack of software facility. The specific area in question here involves data analysis and extraction by a researcher not oriented toward computer analysis.

This thesis describes the design and implementation of a system that allows a researcher with little or no exposure to a computer to use the machine as a tool to help gain insights into the data he is examining. This system was implemented on an IBM System/360 Model 65 computer at the University of Manitoba and an IBM System/370 Model 165 computer at the University of Toronto Computer Centre.

### 1.2 Organization of Thesis

This thesis is divided into three main parts:
(i) an historical review
(ii) the language definition
and (iii) the language implementation

Readers wishing to write programs using the system will be mainly interested in sections (ii) and (iii) corresponding to chapters 3 and $\psi_{0}$ where the language and its implementation is discussed in detail. The user will find a complete program in the appendix, written to show the features of the system.

## HISTORICAL REVIEM

## 2. 1 The Problem

The researcher not familiar with computers faces a difficult task when trying to analyze medium to large amounts of data. Numeric computations can be very time consuming and most researchers prefer to expend their efforts on other aspects of a project. With large amounts of data, the researcher may not have time to do as complete an analysis as he needs to get relevant information from his data base. There are three basic choices he can make.

1) He can try to find which of the many languages in existence is best for his requirements. Then, if the language is available on a machine he has access to, he can proceed to learn the language and do his analysis.

The main fault with this choice is the amount of work that must be done by the researcher before he gets any results. This method does not usually give the researcher any preliminary results that would enable him to decide whether or not his data is significant and whether he should spend the time and money for a more detailed study. There is also difficulty in choosing the proper language. Scme of the widely used general-purpose languages, FORTRAN or $\mathrm{PL} / 1$ for example, will produce the results needed but the researcher will have to learn far
more about the language than he probably needs. If, on the other hand, a more tailored language is found one that specifically suits his needs, the researcher may run the risk of learning a language that he may never have need of again. usually, none of these language types is very useful for a preliminary type of analysis.
2) Researchers in the social science fields often want to count the number of records that correspond to specified conditions. This type of analysis may be done using a card sorter. However, if the number of conditions that are to be met is more than a few, the researcher can be wasting a lot of time getting few results. This method is sufficient for some types of data extraction but requires at least one pass and usually more, over the data for each result required.
3) The researcher's last alternative is to try to $r \in d u c \in$ his data tc a form where it may be manually tabulated. Data or variables thought to be insignificant could be deleted when they are, in fact, significant. Mechanical errors inherent in manual numerical calculations could lead to faulty inferences. It would be a waste not to take advantage of the powerful tools we have access to.

## 2.2 other Solutions

Not very much work has been done to aid the researcher and especially the social science researcher in his efforts to use the computer as a scientific tool. Of what little has been done five languages have been chosen to indicate what is available for the type of research discussed. These languages each show a different philosophy in dealing with the non-computeroriented researcher and the social science researcher wishing to do a preliminary analysis. The five languages are:

1) SRG-40 SURVEY REPORT GENERATOR
2) STATISTICAL REPORT GENERATOR
3) The DATA-TEXT Language
4) COBOL
5) $A P L$
2.2 .1 SRG-40 SURVEY REPORT LANGUAGE

SRG-40 was written for the IBM-7040 computer at the Oniversity of waterloo. Its objective sas to simplify data collection and report writing by computer. Program preparation was to be kept to a minimum without sacrificing efficient use of the computer.

Data input to an SRG-40 program must be in one of two forms, coded or actual. With coded data, a pre-
assigned code is used to indicate an item of data in place of the actual value for example, 1 for male and 2 for female would be adequate codes for a piece of data indicating a certain sex. Actual data indicates that the actual value of an answer is used to represent itself, for example, the number of people over the age of forty in a town or the total salary of a family. Codes can be one or two digits or characters and actual values can be from one to nine numeric characters.

SRG-40s faults lie with its limitations and restrictions. The language has only two basic commands. accumulate (ioe sum) numeric fields and count the occurrences of different coded combinations. This leaves the language uncomplicated for the non-programmer but does not offer him a powerful enough utility. A language should be simple and powerful.

The format of instructions leaves a lot to be desired. Almost everything is defined in terms of numeric codes, making the instructions unintelligible to a person not familiar with the language. This also makes the language harder to learn and remember. Sample commands are as follows:

1) $\mathrm{C} 20104001001002002003003004004 *$
2) A645GR09500*

In this form they are totally meaningless.

The first example would classify part of the data being examined into four categories. The number of times these classifications occurred would then be tabulated.

The "C" indicates we have a code variable. The " 20 " indicates the starting column where the data may be found and the ${ }^{9910}$ indicates length. 0047 tells how many coded groups are defined. The remaining digits define the actual data expected and a numeric equivalent. Each statement is ended by an *。

The second example can be used to sum the values of a specific data field if the data is more than 9500. The "A" defines an actual variable that is to be accumulated. The ${ }^{\text {0 } 64 " ~ i n d i c a t e s ~ t h e ~ d a t a ~ l o c a t i o n ~ a n d ~}$ "5" indicates field width。 "GR" specifies a comparison that is to $b \in$ performed on the data and if the comparison holds true, the accumulation takes place. In this case if the data is greater than "9500" the data will be added into the total. The only comparisons allowed are $G R$ and $L E_{g}$ greater than and less than or equal.

The two statements could be run separately or combined with other statements to count and accumulate based on a number of fields, i.e. count one field and another field AND another and possibly accumulate the values in a number of other fields.

The chairman of the Sociology department at the University of Manitoba was asked to give his impressions of the needs and capabilities of social scientists with respect to computer analysis of their research data. He said that social scientists are not accustomed to computer programming although the situation is improving. When asked what should be included in a language for social scientists, he replied that the statistical capability should be limited, no tests of significance but possibly some histogram capability. He reasoned that there were enough statistical packages available and that the researcher needed a prestatistical capability for counting and tabulating his data.

To this point, the philosophy of SRG-40 agrees, although there is no plotting capability in the language. It is also at this point where the two philosophies diverge. SRG-40 provides no error checking. Program cards must obey format rules and data must be error free. This is contrary to the above belief and the general belief today that languages should be easier to code and provide thorough error checking and good diagnostics. This is especially true for the type of researcher this language is aimed ato

There are a few other faults with SRG-40 that detract from its utility. The format of the report generated by the program is very difficult to read. A
researcher should not have to search for his results When they have been tabulated and are sitting in front of him. Data input to the program must be on cards. Although the majority of social science research data is punched and kept on cards there is no facility available for the times when this is not the case. An even bigger problem is the fact that a user must have one and only one card per record. This means a researcher can only have a maximum of eighty columns of information analyzed at one time.

SRG-40 mas developed around 1965 and it should be kept in mind that it mas uritten for a relarively small machine thus limiting its capabilities.
2.2.2 STATISTICAL REPORT GENERATOR

The second language to be looked at is STATISTICAL REPORT GENERATOR Which will be called SRG (as opposed to SRG-40). Although its name is similar to SURVEY REPORT GENERATOR and they were both developed at the University of Waterloo, there is very little else that is similar.

SRG was designed for researchers needing statistical information and reports but who are not familiar with Data Processing concepts. SRG is a total system, not just a language. In fact, the actual program to analyze data is only one-third of the system. It is this part that is most relevant to the discussion and
will be dealt with in greater detail. SRG moved away from the numeric instructions codes of $S R G-40$ to a more readable English-like format. Instructions are not completely readable but make more sense and are easier to remember. SRG gives users the capability of counting the occurrences of different data combinations as well as summing numeric data items as in $S R G-40$. The user also has the power to define his oun statistical requirements in terms of mathematical expressions. The user has the capability of choosing which data items are to be output in the final report. Sample instructions are as follows:

PRINT NAME AGE SEX
COMPUTE AVG = TOTAL/NUMB
COUNT
IF CONDITION: (1) COUNT

With the last instruction, a count is made only if the previously defined condition, namely condition number $V_{0}$ is satisfied. The condition could be SEX.EQ。 ${ }^{9}{ }^{\text {P }}$ if a count of females is desired or AGE.GT. 40 if a count of people over the age of forty is required. Any instruction except another IF may follow an IF CONDITION. As can be seen. SRG is easier to read and remember than SRG-40.

Like SRG-40, though, SRG has its faults. The language is designed for a terminal system which means
the user must learn the terminal commands and operation as vell as the language. Many installations do not support terminals at all, making it impossible to use. In addition to this, the files that are to be analyzed must be set up off-line. This adds another part of the system to what a user must learn before he can get any results. Altogether, this makes for a fairly complex system, more complex than is needed for a beginning programmer.

The instruction set is fairly easy to use and only has one flaw as far as counting occurrences of different combinations of data goes. If a researcher wants to find the number of occurrences of all permutations of classes of data, he must specify each possible combination in an IF statement. This means a great amount of work for this type of tabulation.

### 2.2.3 DATA-TEXT

Another language that is discussed is DATA-TEXT. This language developed about 1963 at Harvard Oniversity, and since substantially revised (about 1970) for the IBM 360 and CDC 6000 series machines. was designed specifically for data analysis in the social sciences.

The language is very powerful and imposing because of its size. The documentation for the language is so
extensive and voluminous that a beginning programmer has a tendency to feel swamped before he starts．The language offers the researcher almost any and every statistical analysis available．Some of these are listed as follows：

1）Simple statistics－mean。 standard deviation。 variance，count

2）Frequency Distributions
3）Scatterplots
4）Correlations
5）Cross－tabulations and Cross－statistics
6）Factor Analysis and Rotations
7）T－Tests and F－Tests
8）Multiple Regression
9）Analysis of Variance

Data to be input to the language may reside on cards．disk or tape．Variables，representing units of data in a record of a file，are defined in a simple。 readable manner．The user gives each data item a distinct name specifies its location in a record． describes the data item in an English－like manner and describes its codes if it is a coded piece of data．

Actual statistical routines are specified in CCMPUTE statements as follows：

1）＊COAPUTE STATISTICS（AGE，INCOME）
2）＊COMPUTE T－TESTS（ABILITY）

```
*COMPUTE CROSSTABS(RACE&ITEM BY SEX,CLASS)
* COMPUTE REGRESSION (ABILITY ON AGE*SEX INCOME) * GROUP BI CLASS
```

As can be seen, the statistic required is followed by a list of variables to be tabulated along with other needed information. Instructions are readable and fairly easy to remember.

The main fault gith DATA-TEXT is its size. The researcher wishing to do a pre-statistical analysis does not need all that the language has to offer. There is so much available to the user it is difficult to find just What is needed. DATA-TEXT programs can be costly in terms of computer space required. On an IBM machine, DATA-TEXT programs need approximately 200 K bytes before they can run.

DATA-TEXT has been well planned and thoughtfully implemented but it is more suited to the experienced programmer not the researcher unfamiliar sith programming or existing statistical packages.

## 2.2 .4 COBOL

COBOL stands for Common Business Oriented Language. Its chief functions are the maintaining and handing of large complex data files the generation of management reports and numerous accounting applications. It is disliked by many scientific and mathematical programmers
and owes much of its success to the support it received from the Dnited States government, the vorld's largest data processing user.

Programs give the appearance of English and are primarily self-documenting with each program divided into four major sections or divisions. Two of these divisions IDENTIFICATION and ENVIRONMENT, are used for documentation purposes. The DATA DIVISION is used to describe files and variables in a detailed way. The final section the PROCEDURE DIVISION is used to describe the calculations that are to be performed.

COBOL is one of the most widely-used languages in existence today. It is relatively easy to use and provides a variety of facilities for the business programmer. When used in a non-business-oriented application, the language has a few drawbacks. Its selfdocumenting nature requires the programmer to provide information that is not almays required.

Statements are often of a fixed format that must be started within a specific range of columns and must end in a certain fashion. For example the IDENTIFICATION DIVISION must include a PROGRAM-ID starting between columns 8 and 11 followed by a period and a space. The program name then follows enclosed in single quotation marks or apostrophies. This is then also followed by a period and a blank. The language offers such a wide
variety of data types that a user must be aware of not only the type of data he has but in many cases its internal machine representation as well. Many facilities are available to output these data types in a report format. COBOL is hardly every used for mathematical or statistical work. The language has the capability of doing this type of work even though it isp at times, a wk wa rd.

## 2.2 .5 APL

APL is ${ }^{7 A}$ Programming Language" and is the last to be discussed. The language was implemented by IBM for IBM systems back in the $9960^{\circ}$ s and is only now being made available by other companies.

Implementation has been as a remote terminal timesharing interpretive system and thus its use is further restricted to those systems that provide terminal/teleprocessing services.

The last year or two has seen a drive partly spearheaded by Harlan Mills of IBM Corporation, to convert programmers and programming language design to a philosophy called structured programming. Using the proper language and techniques. Mills and others have shown that structured programming dramatically increases programmer productivity as well as the reliability of his work.

APL is the opposite of almost everything structured programming stands for and can be a very difficult language to use or maintain. Variables are defined in much the same way as other languages, that is a combination of letters and/or numbers, starting with a letter. The standard operators are used $t-/ x$ as well as a number of special operators and Greek letters for vector and matrix calculations. When they are all used together it can be very difficult to decipher a program. The following example will help clarify this point. Given a vector of elements $x_{g}$ the line of code shown is sufficient for calculating the average of the elements of the vector.

$$
(+/ x) \div p x
$$

$\rho \times$ gives the number of elements in the vector $x$ $+/ x$ sums the elements in the vector and $(+/ x) \div \rho x$ does the final division to calculate the mean

As problems become more complex. their functional descriptions also become more difficult to understand and read. For the person who does not wish to become bogged down in complex programming problems, apl is not a solution.

The researcher who has large files of data is not given satisfactory service either. At present, there is only one company in Canada, Toronto-based I.P. Sharp. that offers any file-handling capability to speak of. IBM has announced a file handling service for their new
virtual system version of APL but it won't work for almost all of the existing systms that are running today.
In summary, a person wishing to do a short
statistical summary either has to learn too much about
the language or is not offered the necessary facilities
if he wants to use APL.

## DARE = A_DATA REPORT LANGUAGE

### 3.1 Philoscphy

It was felt that the computer languages that existed were not sufficient for solving the type of problem discussed. Throughout the development of DARE (A Data Report languagel every design concept tried to encompass the following principles.

## (1) Simplicity

DARE was designed to be used as a research tool. mainly by social scientists more interested in their data than in "programming". The language has default values for many options and instructions thus helping to cut down on the coding needed. The grammatical structure, or syntax, follows a logical pattern helping to guide the researcher toward his goal. Unlike certain languages such as COBOL; DARE does not require a large amount of coding to do a little work. With DARE, a researcher can do detailed analyses with relatively simple instructions.
(2) Utility

DARE was not designed to replace statistical packages. It was felt that there were more than enough to handle almost any statistical problem a researcher would encounter. Many researchers wanted some type of
pre-statistical analysis, file editing capability and counting facility. DARE gives a user all of these within a variety of formats and options. DARE uas specifically designed for the users who previously spent hours at a card sorter, sorting their data into different classifications or the users who wanted to know if their data had certain preliminary characteristics before they did a lengthy statistical analysis.

## (3) Flexibility

The non-computer-oriented researcher was kept in mind constantly when DARE was being developed. Instructions required only a minimum of information to produce a good deal of analysis. This is fine for many researchers but was still inadequate for those who wanted an improved result. With this in mind, DARE was designed with "options", sections of code that were not necessary for the basic running of the program. but, when used, enabled the user to have a more comprehensive or more readable result.

## (4) Generality

DARE was originally designed to be machine independent. None of the language ${ }^{\text {e }}$ instructions or data types have any bearing or relation to a specific machine or implementation style. Only minor modifications are required to implement the existing version on a wide range of machines.

It was with these four principles that DARE was created. Each served as a guide and played an integral part in the development of the language.

## 3. 2 Lanquage Description

Programs written in the Data Report Language (DARE) have two major sections:
(1) File Definition
and (2) Tabulation and Report writing

### 3.2.1 File Definition

Every file, cr set of data that is to be used, must be defined in the File Definition section. To distinguish one file from another, the files are named in the following manner.

FILENAME (FILEnG)
nn is a two-digit number from 01 to 20. Numbers 0105 have been reserved for input files (files which will be read ing and numbers $06-20$ have been reserved for output files (files which are created by the program). Since most researchers have their data on cards and wish their results on the printer, FILE05 and FILE06 have been reserved for card input and printer output respectively. Users whose data is stored on magnetic tape, disks, or cther devices or who wish to create new files on these devices, define their files in the exact same way as card or printer files.

Up to now, the user has named the file and. in so doing, specified whether it is input or output. He must also describe each field in the file that he wishes to use. Every fiela must have associated with it a distinct variakle name along with other essential information such as its location and data type. This constitutes the minimum the user must declare. In addition to this each variable can have an associated "variable description" for aid in documenting the program and providing a more aesthetic report. Some variable names can also have a "coded description" and this vill be discussed at greater length shortly. The following is the format for defining a variable:

VARIABLE NAME $=$ LOCATION $T Y P E$, VARIABLE DESCRIPTION. (CODED_DESCRIPTION1

Anything underlined is optional.

The different sections are now described.
(1) Variable Name

The variable name is the name a user associates uith one piece of data or one field in his file. It may be a combination of one to six letters and numbers, but must start with a letter. (Some valid variable names are: ADDR, SEX, PROV1, PROV2, P109X. and P74A3.) The following are invalid variable names and the reasons why they are in error are given.

ADDRESS - too long, more than six characters

CODE 1 - contains the invalid character
123HI - starts uith a numeric character

It is suggested that users try to choose meaningful variable names, ones that have some connection with the data they represent. This makes debugging and documentation of the program much easier for the user.

## (2) Location

A field's location is separated from the variable name by an equal sign. This indicates both the end of the variable name and the beginning of the column specification。

A variable's location in the record is indicated by giving the column number that represents its position in the file. An item that covers more than one column is specified by giving the first and last column numbers separated by a hyphen. If an item occupied positions 21 through 25 in a file, these kould be indicated by 21-25. A single column would be represented by a single number. The column location is enclosed in parentheses () and is preceded by the word COLUMN or COLUMNS. The short forms COL or COLS may also be used. The following four examples are all equivalent as far as the definition of the language is concerned.

```
NAME = COLOMNS (10-25)
NAME = COLUMN (10-25)
```

```
NAME = COIS (10-25)
NAME = COL (10-25)
```

(3) Type

The type specification is used to indicate the fundamental characteristics of the data represented by the variable name. In keeping with its objective of simplicity, DARE only allows three data types. alphameric, numeric and coded.

Alphameric variables represent data that can be composed of alphabetic. numeric or special characters. Its main use is for input and output only. Data items that might normally be coded as alphameric (sometimes called alphabetic) are names, addresses, etc.

Numeric variables represent data items that are numbers and nothing else. Generally, the researcher wants some arithmetic operation performed upon these items. Examples of numeric fields could be age, salary, number of children, etc.

The final type specification is coded. This is a special specification that gives DARE most of its power. Often a researcher groups his data into classes according to some criterion. He can then represent each class by its own specific code. For example, a sociologist may be doing a study on the differences in
the cost of living in different regions of Canada. He could subdivide the country into the following regions:

1) Maritimes.
2) Quebec
3) Ontario.
4) Prairies.
5) British Columbia, and
6) North.

The user could then specify the region that a record applied to by specifying the number that was associated with its region or in this case even the first letter of the region's name would suffice as a code. 腰 could stand for the $l$ aritimes. $Q$ for Quebec. etc. Codes can be either alphabetic or numeric and are not restricted to being one letter or digit in size.

The actual type specification is indicated by using the words ALPHAMERIC NOMERIC or CODED after the column specification. $\AA$ few short forms and alternatives have been included to help make the programing and preparation less tedious. aLPHABETIC may be used as an equivalent form of ALPHAMERIC and both may be indicated by the letter $A$. NOMERIC and CODED may be represented by $N$ and $C$ respectively. If the type specification is omitted completely, the variable is assumed to be CoDED.

The type specification can occur immediately following the column specification or they may be separated by either a comma or blanks. It is advisable to use some type of separation for readability.

The following examples show correct ways of defining variables.

1) NAME $=$ COLS (1-15), ALPHABETIC
2) $\operatorname{ADDR}=\operatorname{COLS}(16-30) \mathrm{A}$
3) $\quad$ CITP $=$ COLS (31-45) ALPHAMERIC
4) SALARY $=$ COLOMNS (46-50). NUMERIC
5) $A G E=\operatorname{COL}(51-52) N$
6) $\mathrm{SEX}=\mathrm{COL}$ (55) CODED
7) $\operatorname{PROV}=\operatorname{COL}(56) \quad \ldots(C O D E D$ assumed with no specification)
(4) Variable Description

In addition to the above information a user may add a variable description field. This is a character string enclosed in single quotes ( ${ }^{1}$ ) that helps to describe the variable name. This description is used to aid in the readability of the output results. If the description contains a quotation mark, it should be entered as two single quotation marks. The following could be used as descriptions for the above variable names:

1) ${ }^{\text {(NAME OF FAMILY HEAD }}$
2) ${ }^{\text {a }}$ STREET ADDRESS ${ }^{\circ}$
3) ${ }^{\text {C CITY }} \mathrm{OF}$ RESIDENCE
4) TOTAL INCOME OF FAMIL:
5) FAHILY HEAD' $S$ AGE?
6) $\operatorname{sex}$ OF FAMILY HEAD"
7) PROVINCE OF RESIDENCE ON DECEMBER $31^{\circ}$

These descriptions can immediately follow the type of specification but it is again suggested that the user use either a comma or blanks to separate it from the type specification.

The final specification is the coded description and can only be used with coded variables. It is a means of specifying exactly what each possible code stands for in a coded field. The description should take the form: (CODE/DESCRIPTION,CODE/DESCRIPTION.。.)
with one CODE/DESCRIPTION for every possible code that is used. An example showing the definition of a variable representing a regional code will help make this clearer.

REGION=COLOMN(36)(0/MARITIMES,1/QUEBEC.2/ONTARIO*3/GEST)

Note that the definition here does not need a type specification as CODED is assumed. Alsog there is no variable description but this is not necessary either. The example shows the column specification and coded description separated by a blank. As with the other specifications, this could have been a comma or no space need have been left at all. Each possible code has been listed as a CODE/DESCRIPTION group. This is the best way to code the group even though only the cODE section is necessary. The /DESCRIPTION part helps to document the variatle and makes the output more readable. Some codes are self-documenting and require no description or explanation. For example, the following provincial code will demonstrate this:

$$
\begin{aligned}
& \text { SASK. } \left.A L T A \text {. } B_{0} C_{0}\right)
\end{aligned}
$$

At this point it might be helpful to describe an entire input file every file definition and variable definition must start on a new line. Since card input is the most widely used form of program input, each file definition and variable definition will start on a new card. They may start in any column and continue over to the next if necessary. only card columns 1 to 71 should be used for program statements. If an instruction or definition is too long to fit in the 71 columns. punch any character in column 72 and continue on the next card as if it were an extension of the one being prepared.

Consider a file of student histories containing the studentes name, city of residence, age group, sessional grade and cumulative grade. The data is assumed to be on cards. This file could then be defined as follows:

## FILENAME(FILE05)


CITY $=$ COLS $(16-25)$, $A^{\circ}$ CITY OF RESIDENCE ${ }^{\circ}$

GRADEi=COL (27) :A. SESSIONAL GRADE'
GRADE2 $=$ COI (28), A. ${ }^{\text {P }}$ CUMULATIVE GRADE
Note that the variable definitions are indented from the file definition. This has no effect upon the program and is only used to aid readability. Any file can now be defined using the specifications described above. If a researcher is only in need of a subset of the data he
has in a record, he only needs to define the variables that apply to the information being used. Once the files are defined. the researcher can then move on to the next section, the tabulation section, where he describes what to do with his data.
3.2.2 Tabulation_Section

The tabulation or processing section of a program is the section that specifies what work is to be done: what operations are to be performed.

Eight different operations are available to the user to accomplish his evaluation and preliminary statistical analysis of his data.

The instructions are:

1) READ
2) GRITE
3) MOVE (TRANSFER data or ASSIGN)
4) Calculate
5) CLASSIFY
6) ORDER
7) PLOT
8) IF
3.2.2.1 READ GRITE, DATA TRANSFER

The READ and WRITE instructions will be discussed together because of their great similarity. The format of these instructions is:
instruction (filename)
for example RERD (FILEO5) or
gRITE (FILE 06)

These would cause the information currently stored in the variables associated with FILE05 and FILE06, as specified in a preceding fILE section, to be input and output respectively. File numbers are tested to ensure a READ only accesses an input file and WRITE only accesses an output file. Users are also notified about errors in format and what is expected. The statement

READ) FILE05)
would generate the error message
*** ERROR *** EXPECTING ( BEFORE FILENAME BUT NOT FOUND

Every attempt has been made to make diagnostics as meaningful and helpful as possible. Other error or warning messages associated with input and output specifications are:

## 1) IMPROEER CHARACTER FOLLOGING FILENAME

2) IMPROPER FILE SPECIFICATION FOLLOWING (
3) INCORRECT FILE NUMBER. MUST BE 2 NUMERIC DIGITS
4) FIIE NUMBER OUTSIDE RANGE 1-20
5) NO FILE HAS BEEN SPECIFIED OR FILE DEFINITION MISSING

Because entire files are input and output by a single instruction and file structures are defined with unique variables, there must be a vay of transferring data from one variable to another. A MOVE or assignment statement is used for this transfer and has the following structure.

```
variable1=variable2
```

When this statement is executed the value of variable2 is assigned to variable1. Both variables must $b \in$ of the same type. For coded or alphanumeric data. truncation or padding will take place if necessary at execution.

Extensive error checking is carried out when this statement is translated yielding the following error or tarning messages。

1) a variable in the above statement did not appear IN THE FILE SECTION
2) MIXED MODE IN ASSIGNMENT STATEMENT. VARIABLES DO NOT HAVE THE SAME TYPE
3) TRUNCATION WILL OCCOR AT EXECOTION TIME. CAUSE: FIELD SIZE OF VARIABLES IS NOT THE SAME
4) PADDING RILI TAKE PLACE AT EXECUTION TIME. CAOSE: FIELD SIZE OF VARIABLES IS NOT THE SAME

### 3.2.2.2 CALCULATE

The READ WRITE and MOVE statements are very necessary to any computer language but they are not the focal point of this language. The CALCOLATE statement offers the user the basic functions of summing or totalling fields, calculating averages and standard deviations. The statement takes the form:

CALCULATE operation (OF) VARIABLE1 (BY VARIABLE2)

Farts of the instruction enclosed by parentheses () are optional. Choices for operation are:

1) SUM or TOTAL
2) AVERAGE
3) STANDARD DEVIATION

VARIABLEf must be a numeric variable and VARIABLE2 must be CODED if used.

Sample statements are as follows:

1) CALCOLATE TOTAL PEOPLE
2) CALCULATE SUM OF AGES
3) CALCULATE AVERAGE AGE
4) CALCULATE STANDARD DEVIATION OF ERRORS
5) CALCULATE AVERAGE AGE BY CITY
6) CALCULATE TOTAL WRKDAY BY MONTH

In statements without the BY specification, the numeric operation yields one result, the one that was requested. In the statements with the $B Y$ specification, one numeric result is produced for each of the codes of the coded variable as it uas defined in the file section. In the fifth example above if ten cities had been defined in the file section for the variable CITY, ten averages would be calculated for this statement. each average pertaining to a single city.

Sample diagnostics that can be generated during translation of this statement are:

1) EXPECTING ARITHMETIC FUNCTION AFTER "CALCULATE" BUT IT GAS NOT FOUND
2) EXPECTING VARIABLE NAME OF NUMERIC ITEM AFTER OPERATION BUT NCT FOOND
3) EXPECTING NUMERIC VARIABLE BUT CODED OR ALPHABETIC FOUND
4) EXPECTING "BY" SPECIFICATION OR END OF CARD BUT NON-ALPHABETIC CHARACTER FOUND
5) UNDECODEABLE STATEMENT
6) CODED VARIABLE EXPECTED AFTER "BY" BUT NUMERIC OR ALPHABETIC FOUND
3.2.2.3 CLASSIFY ORDER PLOT

Up to this point the programmer has been able to input and output his data and do the elementary arithmetic functions that make up part of the objectives of this language. The remaining needs of the researcher Within the objectives specified could be satisfied by some type of counting and/or histogram plot facility。

The main method available for counting occurrences of specific data is the CLASSIFY statement. A user can determine the number of times a specific combination of up to twelve variables occurred. Each of the up to twelve variables must be of the CODED type and the instruction yields a matrix-like result. The CLASSIFy statement takes the form

CLASSIFY(variable1,variable2,....variable12)
where all variables are of the CODED type and at least one is listed between the brackets. If $D_{i j}$ represents the jth code of the ith variable of the classification, the results will be listed in the following form:

| D 11 | $D_{21}$ | D $\text { n } 1$ | 等 Of | occurrences |
| :---: | :---: | :---: | :---: | :---: |
| D | D | D | 絡 of | occurrences |
| . 11 | . 21 | - n2 |  | - |
| - | - | - |  | - |
| - | - | - |  | - |
| - | - | - |  | - |
| - | - | - |  | - |
| D | D | D |  | - |
| 11 | . 21 | nam |  | - |
|  | - |  |  | - |
|  | - |  |  | - |
|  | - |  |  | - |
|  | - |  |  | - |
| D | D | D |  | - |
| 1 m | 21 | n 1 |  | - |
| I | D | D |  | - |
| 1m | 21 | n2 |  | - |
| - | - | - |  | - |
| - | - | - |  | - |
| - | - | - |  | - |
| - | - | - |  | - |
| D | D | D |  | - |
| 1m | 21 | nm |  | - |

Every possible combination of all of the codes of all of the variables in the CLASSIFY statement is listed. This statement requires the greatest amount of internal storage of any of the instructions. It was felt that more use would be made of the information that was generated. often a researcher would like to be able to regroup the variables specified in the CLASSIFY statement so that the combinations that result are more functional for his individual requirements. To this end another statement was included. The ORDER statement has the exact same format as the CLASSIFY statement and actually refers to the CLASSIFY statement that precedes it in the program. The ORDER statement contains the same
variables as the previous CLASSIFY but in a different order. An example of these two types of statement is as fcllous:

CLASSIFY $\left(A_{g} B_{g} C\right)$
$\operatorname{ORDER}\left(A_{\theta} C_{\theta} B\right)$
$\operatorname{ORDER}\left(\mathrm{B}_{\theta} \mathrm{A}_{\theta} \mathrm{C}\right)$

The ORDER statement functions in the exact same manner as the CLASSIFY except in one very important respect. The results for the ORDER statement are produced at ${ }^{90} P R I N T$ time and no major storage or execution time is used by the statement. All data that is needed is accumulated by the CLASSIFY statement and When the program is complete, the ORDER statement just re-orders it.

Checking is carried out to insure that all variables used by both statements are CODED and the spncax is correct. Extra checking is carried out in the ORDER statement to insure that the variables listed agree in name and number with the previous CLASSIFY.

Possible error or warning messages issued are:

1) EXPECTING ( AFTER CLASSIFY BUT GAS NOT FOUND
2) MISSING VARIABLE NAME. EXPECTED CODED VARIABLE BUT NOT FOUND
3) TOO MANY VARIABLES IN CLASSIFY INSTRUCTION。

MAX. IS TGELUE
4) VARIABLE USED IN CLASSIFY INSTRUCTION WAS NOT CODED TYPE
 FOUND
6) NO CLASSIFY PRECEDED THE FIRST ORDER STATEMENT
7) EXPECTING (AFTER ORDER BUT NOT FOUND
8) VARIABLE IN ORDER WAS NOT IN PREVIOUS CLASSIFY
9) NUMBER OF VARIABLES IN ORDER STATEMENT IS NOT TGE SAME AS LAST CLASSIFY

The final statement to be described is the PLOT statement. In both of its two forms, the plot statement gives the researcher a histogram plot of either CODED or NOMERIC data. The statement takes the following forms:

1) PLOT BY variable1 where variable1 is a
coded variable
and 2) PLOT variable2 GROUPS=\#1。RANGE=\#2-\#3 where variable2 is a numeric variable.

If the pLOT statement is used with BY and a coded variable, a histogram plot is generated with one bar of the graph representing the occurrences of each code of the coded variable as it was described in the file section.

If the plot statement is used with a numeric variable the user must specify a range or upper and
lower bound for the values which he wants included as the $x$-axis of his plot. If only one value is specified it is assumed to be the upper bound with the lower defaulting at zero. The user must also specify the number of groups (corresponding to bars on his graph) into which the range is to be broken. The y-axis of the graph (i.e. the height of the bars) represents the number of times the value of the variable specified fell into the group plotted.

As with all of the other statements, extensive error checking is carried on with the pLOT statement as well. Error messages genera亡ed are as follows:

1) ILLEGAL CHARACTER SEQUENCE AFTER "PLOT"
2) ILLEGAL CHARACTER SEQUENCE AFTER "BY"
3) ALPHABETIC VARIABLE FOUND IN PLOT STMT
4) EXPECTING "GROUP" OR "RANGE" SPECIFICATION BUT NOT FOUND
5) GROUP SPECIFICATION ALREADY ENCOUNTERED
6) EXPECTING $9=?$ AFTER GROUP BUT NOT FOUND
7) EXPECTING NUMERIC VALUE AFTER GROUPS
8) NO RANGE SPECIFICATION WITH NOMERIC VARIABLE
9) NO GROUP SPECIFICATION WITH NUMERIC VARIABLE
10) EXPECTING NUMERIC RANGE SPECIFICATION BUT NOT FOUND
11) ILLEGAL SEQUENCE FOLLONING RANGE SPECIFICATION
12) TOO MANY NUMERIC BOUNDS FOR RANGE

At this point the user has all of the basic commands needed to do an introductory statistical analysis of his data．He can READ，GRITE，CLASSIFy。 PLOT $_{\text {f }}$ and calculate results in a relatively efficient and easy－to－use manner．It was felt，though，that one additional facility should be added to the language to give it added versatility from the user＇s point of viev． There would always be a time when he would like to apply the existing ccmmands to a specific subset of his data． The best way to do this was to provide a FORTRAN－like IF statement that could be used with all of the commands listed．The IF statement takes the form

IF（logical expression）command

Where command is one of the previously described instructions and the logical expression is a combination of operators and operands that when evaluated yield a result of ${ }^{\text {时true }}$ or ${ }^{\text {＂fals }}$ ．Operands can be variable names，alphabetic or numeric literals．Alphabetic literals are denoted by enclosing them in single quotation marks．operators fall into two main categories，arithmetic and logical．Arithmetic operators consist of + －／and＊for additiong subtraction。 division and multiplication respectively．The logical
 for equal．not equal．greater thang less than．greater
than or equal less than or equal, or and and. All logical operators are both preceded and followed by a period to separate them from operands. Standard FORTRAN priorities are used for operators to help clarify potentially ambiguous situationse Two additional operators are available ( (") and " " ${ }^{\circ 10}$ to help when necessary. The researcher is restricted from using two different types of operands with the same operator. Numeric operands can be used with the following
 example the following statement could be used to calculate the average income of ontario residents whose age is 70 years or more:

IF (PROV.EQ。 ONT0.AND.AGE.GE.70ICALCULATE AVERAGE SALARY
coded and alphabetic operands can use the operators EQ and NE. LOGICAL operators $A N D$ and $O R$ are used to tie parts of the logical expression together coded and alphabetic literals enclosed in quotations should be the exact same length as the other operand they are connected to.

Some of the error messages that can be generated When this statement is scanned are as follows:

1) IF STATEPENT FOUND FOLIONING ANOTHER IF
2) EXPECTING AFTER IF - NOT FOUND
3) IMPROPER OR MISSING OPERATORS/OPERANDS IN IF
4) SYNTAX ERROR - IMPROPER OPERATOR/OPERAND ORDER
5) UNEVEN BRACKETS
6) UNDECODEABLE LOGICAL/COMPARATIVE OPERATOR AFTER -
7) LOGICAL OPERATOR NOT FOLLOWED BY .
8) INVALID LOGICAL OPERATOR FOLLOWING .
9) MIXED MODE - TWO OPERANDS ARE NOT THE SAME TYPE
10) ALPHABETIC OR CODED OPERANDS WITH NON-ALPHA OPERATOR
11) NOMERIC OPERANDS GITH NON-NOMERIC OPERATOR
12) LOGICAL OPERANDS WITH NON-LOGICAL OPERATOR
13) THO CODED VARIABLES AROUND EQ OR NE ARE NOT THE SAME LENGTH
3.2.3 Comment_on_Examples

This now completes the description of all of the instructions available in the tabulation section along with approximately half of the error and warning messages available as diagnostic aids to the researcher. Appendix A contains a sample program that contains deliberate errors to show the error checking capability of the interpreter. Appendix $B$ contains a working program that shows some of the capabilities of the language. The problem solved by the program is a very common one for almost all computer centres, a routine analysis of accounting data produced by programs run on a typical day. The data used pertains to the System/370 installation at the 0 niversity of Toronto. The centres Oser Services staff feels it needs to know the distribution of jobs at each of the more than ten remote
terminals that feed the main system. They need not only the number of jobs from each remote but specific information about the individual jobs as well. The system measurement group also keeps records on resource utilization. With the program in Appendix $B_{0}$ all the information that is desired is available。

One final item should be noted with regard to the results from the tabulation section. All results are output automatically in a fixed format. The nature of the language and its objectives does not lend itself to burdening the researcher with having to specify elaborate formatting schemes. Each line of a program is listed by the interpreter with a line number. All results generated by an instruction in the tabulation section refer back to the line number of the instruction that generated it. Although this is not the most elegant method it is functional and easy to use.

## 4. 1 HEDABIN - Manitoba_Data_Report_Interpreter

4.1.1 ㄴhilosochy of Implementation

A great deal of thought went into the decisions on how to implement the language defined. The major decision to be made was whether to implement it as a compiler-based system or an interpretive system. There were definite advantages to both.

A compiler-based system was better for applications that would be performed repetetively. A load module could be produced once and this module could then be used to do the analysis. The compiler-based system was more efficient There was no scanning needed once an object module was produced and the object module itself would be in machine or assembly code.

An interpretive system had other advantages. Users. especially nowice ones. were not bothered with object modules and load modules. A system written in a highleyel language that execured its otn code was one step closer to being machine independent. If it was anticipated that the language would not be used for a production-like purpose then the overhead incurred could easily be balanced by the Elexibility gained.

The solution that was eventually arrived at involved a combination of the advantages of both the
compiler-based and interpretive systems. All programs are scanned for correct spntax and an internal code is produced for the interpreter. It is this simplified code that is executed, eliminating many of the problems and much of the overhead of conventional systems. Both the translator and interpreter have been uritten in a highlevel language and although not totally machine independent are as close as can be。
4.1.2 Gross_Logic

The system is designed to be a batch processing system, that is many jobs could be processed and run with one loading of the inter preter.

Every job goes through four phases while being processed. The phases are individual processors that are necessary for translating the user s program into a form that is needed to manipulate his data.

A CONTROL PROCESSOR is used to act as a supervisor passing control between the other processors. The FILE PROCESSOR handles all set-up requirements for $I / O$ and the TABULATOR PROCESSOR looks after executable statements. Once all of the statements have been translated, control passes to the EXECOTION PROCESSOR for execution.

These are now discussed in greater detail.
4.1.3 MERDARIN ROUTINES
4.1.3.1 CONTROL PROCESSOR

A control section was developed to differentiate between different jobs of the same batch as well as sections of the same job. Control cards were implemented that indicated the different stages of processing that a user would go through. The control cards are:

1) JOB
2) FILE
3) TABOL酚E
4) DATA
5) EJECT
and 6) COMMENT

Two other control cards were added as aids to debugging the interpreter:

1) DEBUGON
and 2) DEBUGOFF

Every control card must be preceded by a control character in the first column of the card. This is a reserved character and may not be used for any other purpose if it is in column 1. All control cards take the form
control character control statement

In the current version of the interpreter 0 is used as the control character. Appendix A and B contain examples of its use. The comment statement is slightly different. Instead of specifying

ØCOMENT
for a comment card, the user only has to specify two control characters. ๑๐.

Each control card has a very definite function. The JOB card is used to separate jobs in a batch as well as pass additional information on to the interpreter. When encountered the CONTROL PROCESSOR passes control to routines that clean up any previous job in the same batch another routine is then performed that initializes data for the new job and picks up job options from the job card. These include the ability to suppress warning messages from the program listing and request the use of 2 byte integers for space saving in memory. The user requests these options by specifying keywords anywhere after the word JOB. The keywords for suppression of warning messages and the use of a smaller word size are NOWARN and SMALL respectively. The useris name and/or identification, if enclosed in single quotation marks, is also picked up from this card and is used when a job separator page is produced. Following the production of the separator page and variable initializationg control returns to a main input routine to get the next card.

The FILE card is usually found following the job card. When encountered the CONTROL PROCESSOR initializes variables needed in the definition of $I / O$ files and returns to the main input routine.

The TABULATE card invokes routines that terminate the previous processor. Variables needed for the TABULATE PROCESSOR are initialized and control returns again to the main input routine.

The DATA card signifies that all commands or instructions have been entered. Internal code is generated to provide a "flow control" through the code that was previously generated. Variables needed in execution are initialized and if no errors have been recorded to this point, control passes to the EXECUTION PROCESSOR.

The fifth control card is the EJECT card. This is only used to aid in the ouput listing of a program. When encountered, the printer is forced to skip to the top of a new page where the listing is continued.

The COMMENT card is a convenient method of supplying comments or documentation in a program. All comments are listed in the order they are found and have no other effect upon a program.

The DEBUGON and DEBJGOFF cards are used to selectively turn on and off special procedures used to
debug the interpreter. They are of very limited use to the programmer but can be of significant help to the systems personnel responsible for the interpreter. The backbone of the debugging procedures is a formatted dump routine that prints out main memory for the interpreter along with some of the important pointers and variables.

In addition to providing the control between each of these control sections, the CONTROL PROCESSOR does a certain amount of error checking to insure that program sections are in the proper sequence and one program does not destroy another one in the same batch.

### 4.1.3.2 FILE PROCESSOR

A file section handles the formation of data structures needed by the interpreter as well as the allocation of storage space for variables defined by the user.

Control is passed from the main input routine to the analysis portion of the file section. If the card just input indicates the beginning of a new file fiee. a FILENAME card) the last file is closed off and the new file is stored in the interpreters memory. All files are linked together to provide the capability for efficient file searches. Checks are provided to insure that the user does not try to define the same file more than once.

If the card passed to the file section is not a FILENAME card it is assumed to be a variable definition card. Routines are called that break this input card into syntactic units. If the syntax of the statement is correct the data that has been broken down is used to build the information about the variable that will be needed at execution time. The FILE PROCESSOR links all of the variables of a program together for efficient variable searches. As uith the file definitions, checks are made to insure unique variable names in the definition phase。

### 4.1.3.3 TABOLATE PROCESSOR

The third section is the tabulation section. Its function is to do all syntax checking of executable statements, issue error and warning messages when necessary, and generate the internal text for the interpreter.

In many respects, the TABULATE PROCESSOR is similar to the FILE PROCESSOR. Statements passed to the processor have their first syntactic unit used in a table look-up sequence. If the unit matches any element in the table, control is passed to a specialized routine that continues syntax checking and internal code generation. The specialized routines use the same code as the FILE PROCESSOR for doing the syntax analysis and error message generation. If a statement is
syntactically correct, its equivalent internal code is loaded into memory. All variable references are resolved and run-time error checking is included if needed. Control returns to the main input routine so that the next statement may be read in. The tabulate processor also keeps track of the memory location of the first statement encountered and this is passed on to the EXECUIION PROCESSOR.

### 4.1.3.4 EXECUTION PROCESSOR

The final processor is the interpreter. It is responsiblefor executing the code produced by the TABOLATE PROCESSOR and for issuing execution-time error messages. The interpreter also keeps track of certain job statistics such as memory utilization, the number of records read from an input file or written to an output file, and the number of error and warning messages generated for the job.

The EXECUTION PROCESSOR uses three main parts of the interpreter, an Instruction Address Register, a branch table and instruction modules. The Instruction Address Register (IAR) is initially set to point to the first instruction in memory. This is passed to the EXECUTION PROCESSOR from the TABULATE PROCESSOR. The EXECOTION PROCESSOR picks up the byte in memory pointed to by the $I A R$ and uses this as an index into a branch table. The byte of information actually represents an
operation code of a specific instruction. The branch table contains pointers to the different routines needed to execute the different operation codes. When control passes to these routines, they use the information following the operation code as operands. After processing the operation code, the IAR is incremented by a fixed number of bytes so that it points to the next operation code. Control passes back to the point where the next operation code is picked up and used in the branch table.

Each instruction (operation code) has two phases. READ and WRITE instructions are executed completely each time they are encountered in a looping program. This is what one would normally expect. A statement requesting the $A V E R A G E$ value of a field cannot be fully executed until all of the data has been collected. The two phases for each instruction could be called the "loop phase and the "final output phase"。 For READ and VRITE instructions, a READ or WRITE is performed every time it is encountered in the loop phase. CALCULATE instructions only collected data during the mloop phase". The loop phase ends and the "final output phase" begins when all the data has been collected, that is when an end-of-file interrupt is raised while doing a READ operation. At this point one final circuit is made through the program not to add new data to totals or averages but
to complete the final calculations and output the results. This is the "final output phase".

### 4.2 Internal Organization

### 4.2.1 Symbol Table Entries

The symbcl table is located at the low end of the interpreter ${ }^{9}$ memory. It contains the internal format of all of the files defined in the tabulation section of a program. All variables defined are entered into the table and are connected to all previously defined variables in a linked list structure. This allows for efficient searches of the symbol table while allowing for variable length entries in the table. In addition to a link each entry in the table contains the variable name, and displacement from the beginning of the record for which the data item is defined, the length of the data item, its type (numeric, coded or alphanumeric), a pointer to a memory location where the description of the variable is located if the description has been included and a pointer to the coded description if the variable is of the coded type. Additional space has been included for possible future expansion but is not presently used.

### 4.2.2 Instruction Entries

The following sections describe briefly the internal code that is generated by the different processors mentioned above. Each statement from a
program is broken down into an operation code and operands.

$$
4.2 .2 .1 \text { READ }
$$

The READ instruction takes the following format: OP CODE, FILE-NUMBER-POINTER

The File-Number-Pointer points to a memory location where the file structure for the file in question is located.
4.2.2.2 VRITE

The GRITE instruction takes the identical format to that of the READ.

OP CODE FILE-NUMBER-POINTER
The pointer indicates the area defining the file structure and data to be output.
4.2.2.3 MOVE

The MOVE instruction has the form

$$
\text { OP CODE, POINTER } 1, \text { POINTER } 2 \text {, Length }
$$ padding (if necessary)

Pointer 2 points to the information to be moved while Pointer 1 points to the area into which the data is to move. Length represents the number of characters to be transferred and padding represents the number of blanks to be used to fill out a field that is larger than need be.

## 4.2 .2 .4 CALCULATE

The CALCULATE statement is really three separate statements, one for SUM, AVERAGE and STANDARD DEVIATION. Each consists of the following:

OP CODE, CODE-COUNTER。DATA-POINTERS。 ACCOHOLATORS

The code-counter is used when a calculate statement is used with a "BY CODED variable phrase as in the following:

Calculate average age by city
Where CITY is coded.

The code-counter contains the dimensions of the variable CITY di.e. the number of codes that were defined for that variable). The data-pointers point to the memory locations of the CODED variable and the data to be used in the calculations. The value of the codecounter specifies the number of sets of accumulators needed for the computation. If, in the above example, seven cities mere defined, we yould need seven sets of accumulators, cne for each city. The following table indicates the number of accumulators in each set for each CALCOLATE statement:

Statement $\quad$ \# of accumulators
SOM/TOTAL 1

AVERAGE
2
STANDARD DEVIATION
If a CODED variable is not used, one set of accumulators is sufficient for all calculations.

$$
4.2 .2 .5 \quad \text { ISN }
$$

The ISN instruction is generated at the beginning of every program statement. It takes the form OP CODE I INE-NOMBER

The line-number refers to the statement number of the statement about to be processed. hhen encountered during execution this statement updates an instruction counter that can be used in execution-time error messages and diagnostics.

### 4.2.2.6 CLASSIFY/ORDER

CLASSIFY and ORDER are treated in the same section because of the similarity of their organization and the way in which they are tied together. The CLASSIFY instruction takes the following form:

OP CODE $V$ VARIABLE-COUNT, VECTOR-SIZE VARIABLEPOINTERS*DIMENSIONS MULTIPLIERS*COUNTER-VECTOR

VARIABLE-CCUNT (n) is the number of variables in the statement. VECTOR-SIZE (m) is the product of the dimensions of the $n$ variables. VARIABLE-POINTERS point to the memory locations of the $n$ variables. DIMENSIONS represents the $n$ dimensions of the $n$ variables, and MUITIPLIERS ( $n$ of them) are used for indexing into the vector of counters and are defined as follows:

COONTERS (m of them) are used to count the occurrence at each possible combination of variables.

The ORDER statement has the following form:
OP CODE, VARIABLE-POINTERS,DIMENSIONS。 MOLTIPLIERS

All of these have the same meaning as the CLASSIFY with cne exception. If variable $i$ in the order statement has the position $j$ in the previous CLASSIFY multiplier $i$ in the ORDER instruction takes the value of multiplier $j$ from the previous CLASSIFY instruction. When these multipliers are used to index into the vector from the previous CLASSIFq they will access the same data in a new order.

$$
4.2 .2 .7 \quad \mathrm{IF}
$$

The IF instruction follows a very simple logic and takes the form:

OP CODE ERPRESSION-RESULT-POINTER*NEW-INSTRUCTION-ADDRESS

The Expression-Result-Pointer points to an area of memory that contains a true or false value ( 1 or 0 ) representing the validity of the logical expression originally specified in the IF statement of the program. If the statement was true, the Instruction Address Register (IAR) is set to point to the next instruction in memory. If the statement was false, the $I A R$ is assigned the New-Instruction-Address from the IF instruction. This new address is just the location
after the instruction that followed at the end of the IF statement in the program.

$$
4.2 .2 .8 \text { PLOT }
$$

There are tho forms of the plot statement, one for CODED data and one for NUMERIC.

The CODED PLOT takes the following form:

OP CODE, DIMENSION $V$ VARIABLE-ADDRESS_COUNTERS

DIMENSION(n) is the number of entries for the CODED variable. VARIAELE-ADDRESS points to the memory location defining the variable being plotted. COUNTERS represents the $n$ accumulators needed for the $n$ bars of the histogram plot.

The NOMERIC PLOT takes the following form:

OP CODE, GROUPS, VARIABLE-ADDRESS, RANGE-1。 RANGE-2。 COUNTERS

GROUPS represents the number of segments that are to be plotted in the numeric span between RANGE-1 and RANGE-2. VARIABLE-ADDRESS points to the variable definition of the data being plotted and counters represents the $n$ accumulators needed for the $n$ bars of the histogram plot.

Each of the above instructions was designed to use as little memory space as possible but still provide the necessary information for efficient execution.

## Conclusion

In trying to evaluate this thesis the author has asked four basic questions:

1) Does the implementation of the language follow the design specifications and work as it should?
2) Does the project meet the objectives originally proposed?
3) With the aid of hindsight, how would the project have been changed to possibly attain greater results?
4) Vas the project a success?

The implementation of the language follows the specifications exactly and does indeed work. Extensive testing was performed and the interpreter has no known "bugs". By far, the largest number of problems that arose during final testing of the interpreter were due to keypunch errors in the programs being tested or the data being used. These were all caught by the interpreter and the diagnostics that were generated made the error correction trivial.

There were four basic objectives originally considered and most of them have been met. Simplicity has been maintained throughout. The language is restricted to a fev basic. powerful commands that are easy to use but still do the mork. The language seems to have a wide range of applications even in a production environment for daily statistical applications as can be
seen in the example in the appendix. Flexibility has been maintained from the language definitions through to the interpreter implementations. Options have been provided to make the use of the language as simple as possible. The final objective, generality, has not been completely attained. Even though the interpreter was uritten in a high-level language a few sections of code were written in a machine-dependent manner. These sections are very few in number and should not produce many problems if a version is required for another machine (i.e. other than IBM/360 or 370 ).

Two design decisions would be different if the project was being re-written. one would be to write specific parts of the translator in a low-level language for greater efficiency. The second would be to write a compiler version instead of an interpreter. Originally When this language was still in the design stage it was felt that its main uses would be as a pre-statistical package data analyzer. A researcher with his data would run a program a few times to determine whether or not to go on and do an in-depth analysis. Since the interpreter has been finished more and more applications have developed that can use the language as it is and do not need anything else. For these applications that are run on a regular basis, an interpreter does not offer the efficiency that is desired and can only be accomodated by a load module。

The project can definitely be termed a success for the fcllowing two reasons:

1) The project met or exceeded almost every objective.
2) The author has gained a great deal of knowledge in the areas of language design. language implementation and large-project implementation。

APPENDIX_A

This section contains two sample jobs with deliberate programming errors. It is meant to show the diagnostic features of the interpreter.




トトトトトートトトトトトトトトトートトトトトトトトートトトトトトトト












MANITORA DATA REPDRT INTEFPRETFR

品品号


站记





NNNNNNNNNNNNNNNNNNNNNNのNNNのNN


结过 0

 $\frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{s} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{4} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{3} \frac{\alpha}{4} \frac{\alpha}{3}$

SMALI






 언

 ভ之




 MANDAR
$M A N D \triangle R$
$M A N O A R$
$M A N D A R$
$M A N D A R I$
$M A N D A R$
$M A N D A R$
$M A N \cap A R$
$M A N D A R$
$M A N O A R$
$M A N D A R$
$M A N D A R$
$M A N D A R$
$M A N D A R$
$M A N D A R$
$M A N D A R$
$M A N D A R$
$M A N O A R I$
$M A N D A R I$

문운

$$
\alpha \alpha
$$



APPENDIX_B

This section contains a sample working job. It not only shous that the interpreter works but also shous scme of its functional capability. The program shown solves a problem many compurer centres have. Given accounting records from jobs rung do a statistical analysis on certain aspects of the data. The oniversity of Toronto has a very large centre with many small terminals scattered throughout the different campuses that act as feeders to the central site. Management likes to keep a profile of the types of jobs that run on the system, that is, where they come from what priority do they use, cards read ing lines printed, etc. The file section of the example defines the structure of one of the accounting records. The tabulate section contains the instructions for the analysis. The first three instructions generate plotted output. The next eight instructions calculate required statistics and yield a numerical output. The CLASSIFY and ORDER statements generate the bulk of the output and count the occurrences of the different combinations of job options available at the centre. The final statement counts the number of forms meunts of jobs run priority 8 and submitted through the terminal system.


[^0][^1]



SMALL



MINIMUM
0
COCE/DESCR•OR RANGE
CDCE/DESCR。OR RAN
HASP/OS.JOB CLASS






















































## BIBLIOGRAPHY

1．APL／360 Primer．IBM Form Number C20－1702．
2．API－PLOS－File－Subsystem Instruction Manual。 Toronto： I．P．Sharp Associates Itd． 1972 。

3．Armor．D．J．＂A Computer Language for the Analysis of Variance，＂1 reprinted from the 1969 Social Statistics Section of the American Statistical Association。

4．Armor，D．J．and Couch．A．S．Data Text－primer（final draft of the book published by The Free Press．New York 1972）．

5．Bennett．Rok．＂The Design of Computer Languages and Softuare Systems：A Basic Approach。p Computers and Automation（February 1969）．
 report of the University of Naterloo．April 1969．

7．Cress．P．H．and Dirksen．P．H．＂SRG－40－Survey Report Generator．＂A report of the 0niversity of Waterloo， 1968 。

8．Cress．P．H．et al．＂Description of 1360 WATFOR．A Fortran IV Compiler．A Report of the Faculty of Mathematics．University of Waterloo，1968．

9．Fortran－IV－（G）－Prodram Loqic Manual．IBM Form Number Y 28－6638。

10．Fortean－IV－（H）－Program－Logic Manual．IBM Form Number Y 28－0012．

11．Full－American National－Standard Cobol．IBM Form Number GC28－6396．

12．Genuys．F．（ed．）Programming－Lanquages－Nato Adyanced Study－Institute．New York：Academic Press．1968．

13．Rosen．S．＂Programming Systems and Languages－A Historical Survey．＂in Programming Systems and Lanquages．ed．S．Rosen．New York：McGraw－Hill． 1967.

14．Sammet，J．Programming－Lanquages：History and Fundamentals．Englewood Cliffs，New Jersey： Prentice Hall．1969．
15. Shantz, P. W. et-al. "Watror - The University of Waterloo Fortran IV COMPILER ${ }^{13}$ Communications of the ACM, $X$. No. 1 (January 1967).


[^0]:    씄

[^1]: