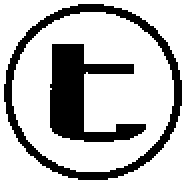


S))))))))))))))))))))))))))))))))))

23A Durham Drive
Amherst, NY 14228

Galileo*CATPAC:

User Manual and Tutorial

Rev. June, 1993

S))))))))))))))))))Q

CATPAC COPYRIGHT 1990 BY JOSEPH WOELFEL

PLEASE DON'T LOSE THIS PAGE. IT CONTAINS THE REGISTRATION NUMBER YOU WILL NEED TO UPGRADE TO LATER RELEASES OF CATPAC.

Your Name _____
Your Registration Number _____

Version 3.0

Terra Research

IMPORTANT!
PLEASE READ CAREFULLY BEFORE USING THE SOFTWARE.

NOTIFICATION OF COPYRIGHT

THIS SOFTWARE IS A PROPRIETARY PRODUCT OF The Galileo Company AND IS PROTECTED BY COPYRIGHT LAWS AND INTERNATIONAL TREATY. YOU MAY MAKE A REASONABLE NUMBER OF COPIES OF THIS PROGRAM FOR BACKUP PURPOSES, AND YOU MAY COPY THE SOFTWARE TO THE HARD DISK OF A SINGLE COMPUTING PLATFORM OF THE TYPE SPECIFIED IN YOUR LICENSE.

YOU ARE PROHIBITED FROM MAKING ANY OTHER COPIES OF THE SOFTWARE FOR ANY OTHER PURPOSE BY COPYRIGHT LAWS. YOU MAY MAKE ONE COPY OF THE WRITTEN MATERIALS ACCOMPANYING THIS SOFTWARE FOR ARCHIVAL PURPOSES.

The Galileo Company

PLEASE READ THIS LICENSE AGREEMENT BEFORE USING THE SOFTWARE. THIS AGREEMENT IS A LEGAL CONTRACT BETWEEN YOU AND The Galileo Company GOVERNING YOUR USE OF THIS SOFTWARE. USING THIS SOFTWARE INDICATES YOUR ACCEPTANCE OF THIS AGREEMENT. IF YOU DO NOT WISH TO ACCEPT THE TERMS OF THIS AGREEMENT, PLEASE RETURN THE UNOPENED SOFTWARE PROMPTLY TO The Galileo Company. IF YOU HAVE ANY QUESTIONS ABOUT THIS AGREEMENT, PLEASE CONTACT The Galileo Company, 23A Durham Drive, Amherst, NY, 14228.

TERMS OF LICENSE

THIS IS AN EXPERIMENTAL PROGRAM. WHILE The Galileo Company CERTIFIES THAT THE HIGHEST STANDARDS OF DILIGENCE AND SCIENTIFIC INTEGRITY HAVE BEEN APPLIED TO THE DEVELOPMENT OF THIS SOFTWARE, BY ACCEPTING THIS LICENSE YOU AGREE THAT THIS IS EXPERIMENTAL SOFTWARE AT THE CUTTING EDGE OF SCIENTIFIC PROGRESS.

NOT AS MUCH IS KNOWN ABOUT THE PERFORMANCE OF NEURAL NETWORK TECHNOLOGY AS IS KNOWN ABOUT TRADITIONAL COMPUTER SOFTWARE. YOU AS THE END USER AGREE THAT REASONABLE AND PRUDENT CAUTION ABOUT THE APPLICATION OF RESULTS FROM THIS SOFTWARE IS APPROPRIATE, AND The Galileo Company AGREES TO SHARE WITH YOU (THE LICENSEE) RELIABLE ESTIMATES OF THE OPERATING PARAMETERS OF THE SOFTWARE INSOFAR AS THEY ARE KNOWN BY TERRA.

The Galileo Company GRANTS YOU THE RIGHT TO USE ONE COPY OF THE SOFTWARE ON A SINGLE-USER COMPUTER. EACH WORKSTATION OR TERMINAL ON A MULTI-USER COMPUTER SYSTEM OR LOCAL AREA NETWORK MUST BE LICENSED SEPARATELY BY TERRA RESEARCH AND COMPUTING COMPANY.

YOU MAY NOT SUBLICENSE, RENT OR LEASE THE SOFTWARE TO ANY OTHER PARTY.

YOU MAY MAKE REASONABLE BACKUP OR ARCHIVAL COPIES IF THE SOFTWARE, BUT YOU MAY NOT DISASSEMBLE, DECOMPILE, COPY, TRANSFER, REVERSE ENGINEER OR OTHERWISE USE THE SOFTWARE EXCEPT AS STATED IN THIS AGREEMENT.

LIMITED WARRANTY

The Galileo Company will replace defective diskettes that are returned within 90 days of the original purchase date without charge. The Galileo Company warrants that the software will perform substantially as stated in the accompanying written materials. If you should discover any significant defect and report it to The Galileo Company within 90 days of purchase, and Terra is unable to correct it within 90 days of receipt of your report of the defect, you may return the software and Terra will refund the price of purchase.

SUCH WARRANTIES ARE IN LIEU OF OTHER WARRANTIES, EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE WITH RESPECT TO THE SOFTWARE AND THE ACCOMPANYING WRITTEN MATERIALS. IN NO EVENT WILL The Galileo Company BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY LOSS OF PROFITS, LOST SAVINGS, OR OTHER INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF YOUR USE OF OR INABILITY TO USE THE PROGRAM, EVEN IF The Galileo Company OR AN AUTHORIZED TERRA REPRESENTATIVE HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. The Galileo Company WILL NOT BE LIABLE FOR ANY SUCH CLAIM BY ANY OTHER PARTY.

This limited warranty gives you specific legal rights. Some states provide other rights, and some states do not allow limiting implied warranties or limiting liability for incidental or consequential damages. For this reason, the above limitations and/or exclusions may not apply to you. If any provision of this agreement shall be unlawful, void or for any reason unenforceable, then that provision shall be deemed separable from this agreement and shall not affect the validity and enforceability of the remaining provisions of this agreement. This agreement is governed by the laws of the State of New

York.

U.S. Government Restricted Rights

The software and accompanying materials are provided with Restricted Rights. Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (b)(3)(ii) of the Rights in Technical Data and Computer Software clause at 252.277=7013. Contractor/manufacturer is The Galileo Company,23A Durham Drive, Amherst, NY 14228.

TABLE OF CONTENTS

INTRODUCTION TO CATPAC

CATPAC™ is a self-organizing Artificial Neural Network that has been optimized for reading text. **CATPAC** is able to identify the most important words in a text and determine their patterns of similarity based on their associations in the text. From this information, it is able to tell you the main concepts dealt with in the text.

CATPAC does this by assigning a neuron to each major word in the text. It then runs a scanning window through the text. The neuron representing a word becomes active when that word appears in the window, and remains active as long as the word remains in the window. Up to *N* words can be in the window at once, where *N* is a parameter set by the user.

As in the human brain, the connections between neurons that are simultaneously active are strengthened following the law of classical conditioning. The pattern of weights or connections among neurons forms a representation within **CATPAC** of the associations among the words in the text. This pattern of weights represents complete information about the similarities among all the words in the text.

Technically, the pattern of connections among neurons is a complete paired comparison similarities matrix, and so lends itself to the most powerful and sophisticated of statistical analyses. Among these is the diameter method cluster analysis automatically performed by **CATPAC**.

CATPAC can automatically exclude from consideration any arbitrary list of words. A default list of articles, prepositions and the like, is contained in a file labeled EXCLUDE.DAT. You can add or delete words to this file by using the enclosed program EXCLUDE.EXE

CATPAC expects to find both EXCLUDE.DAT and EXCLUDE.EXE in a directory called C:\GALILEO\RUNNER. This directory is automatically created during installation. If you have not followed the standard installation instructions, you must create such a directory and copy all files with an .EXE extension, as well as the EXCLUDE.DAT file, to this directory.

You can create you own exclude files and we recommend that you do so. Every data set is different and there are some words you may wish to exclude in one but not another. To use an exclude file you have created, simply enter it's name in the Exclude file field.

We also recommend that you place the directory C:\GALILEO\RUNNER in your path. This will allow you to run CATPAC from any directory, simplify your analysis set-ups, and keep your raw data (and/or results) separate from your program files. If you are not sure how to edit the path statement contained in your AUTOEXEC.BAT file, consult your DOS manual.

INSTALLING CATPAC

- Place the diskette in the A: or B: drive.

- Type **INSTALL <diskette drive> <target drive>** and press Enter.

  For example to install the system on your C: drive with the diskette in the A: drive you would type:

  INSTALL A: C:

That's it. The install program will take care of everything.

The following directories will be created:

\GALILEO\RUNNER	Contains the executable programs
\GALILEO\HELP	Contains the help files
\GALILEO\DOC	Contains all available Galileo Documentation in WordPerfect 5.0 format
\GALILEO\DATA	Contains sample data sets
\GALILEO\TOOLS	Contains a text editor and several utility programs

RUNNING CATPAC

You will need at least 530k of free ram to run Catpac. If you experience difficulty running the program, chances are you will have to free up some memory. The best way to rectify a memory problem (if you have a 386 or better processor) is to install a memory manager like QEMM. This program is inexpensive and will give you up to 630k of ram to run programs with. Another approach would be to create a 'vanilla' boot diskette that has the bare minimum of TSR's and device drivers in Autoexec.bat and Config.sys. If you are unfamiliar with these files or are unsure how to make a boot diskette, refer to your DOS manual or resident Techie.

There are 3 ways to access **CATPAC**:

(1) Type the word **GALILEO** and a menu will appear on the screen. You would then select the number which corresponds to **CATPAC**. If you have not edited your path to include **\GALILEO\RUNNER**, you must first change to C:\GALILEO\RUNNER prior to typing **GALILEO**.

(2) If you have placed C:\GALILEO\RUNNER in your path, you can simply type **CATPAC**

(3) If you have not edited your path, you must first change to the directory C:\GALILEO\RUNNER and then type **CATPAC**

If you have not installed a memory manager or have a 286 machine and are having difficulty running the program try either method 2 or 3. These methods require less memory because they do not use the menu interface.

Once you call up the program, **CATPAC** will display a menu screen like the one below. To make a selection, press the function key that corresponds to the operation you wish to preform. You can choose one of three types of analysis, Help, or Reset to defaults. Each of these options is explained below.

2 CATPAC main menu

SCREEN CONTROL

Each analysis screen has a number of fields that require information to run (see figure 3). You move from field to field using the arrow or tab keys.

Once you have entered all the information, press **F10** to run the job. If you wish to bail out and go back to the Main Menu, press **F7**.

One nice feature of **CATPAC** is the ability to do multiple runs with the same settings. If you decide to analyze 10 data sets each with the same settings you need enter the settings but once. You need only change the Data file and File descriptor field (or output file fields for a Frequency or Cluster analysis).

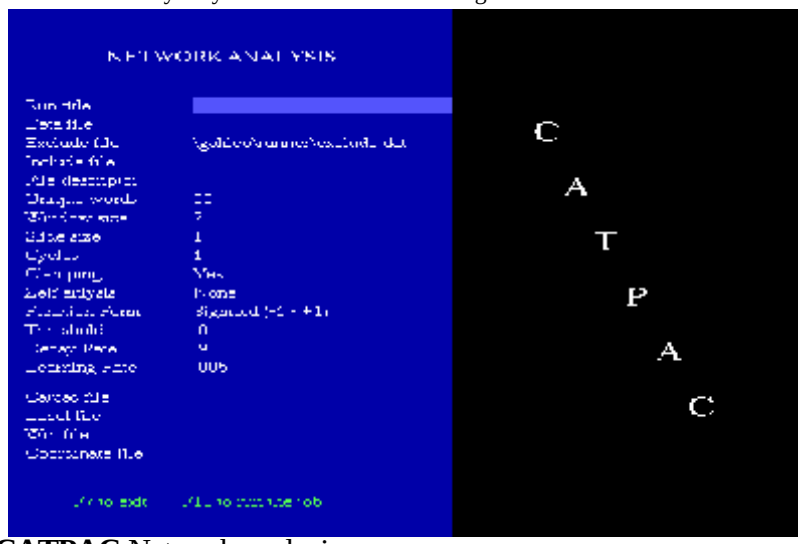
NETWORK ANALYSIS

This is the premier analysis offered by **CATPAC**, accept no substitutes. If you want to perform a Cluster analysis based upon simple word co-occurrences, you would have typed **F2.** Essentially, this was way all cluster analyses were performed before the development of artificial neural networks like **CATPAC**.

You, however, choose to conduct a more advanced form of relational analysis, a Network Analysis, so you pressed **F1.** In a Network analysis CATPAC will generate square matrix of numbers which summarizes the connection strength between unique words. How catpac performs this analysis is discussed in detail below as we explain the fields specific to a Network analysis.

> **Be aware that if you want to produce a perceptual (or brand) map, you must choose a Network analysis.**

If you Press **F1** to do a Network analysis you will see the following screen:



3 CATPAC Network analysis screen

Some of the fields for a Network analysis are also offered for a Cluster analysis and a Frequency run so we'll explain them here. Many of the fields are supplied with default values and file names that we have found to work best with a wide variety of runs. You are encouraged to change them at your whim to experiment with different settings. Every data set is different and what works for one may not work for another.

Run title

CATPAC is asking you for a title, which will be printed as a banner on your output file. Your title is limited to forty characters (including spaces).

Data file

CATPAC needs to know the exact path to your data. You must specify: the drive, the directory and the name of the file which contains the text you wish to analyze. If you are running **CATPAC** from the directory where your data are stored, you can simply type the name of the file which contains your

data.

Remember, **CATPAC** can only read an ASCII file. Sometimes, people make the error of exiting their word processor without (first) converting the file to ASCII. Don't make this mistake!

We like to use the extension .TXT to denote an ASCII file. That way, when we see this extension on a file, we know **CATPAC** can read it.

Exclude file

CATPAC will let you use any EXCLUDE file you wish. If the one we sent you (EXCLUDE.DAT) meets your needs, simply skip this field to accept the default EXCLUDE file. Alternatively, if you want to use another EXCLUDE file you have created for a special purpose, type the name of this file. Again, if this file is not in your current directory, be sure to specify its exact path. Make sure you Exclude file is in ASCII and has only one work on each line.

Include file

Just as you can exclude certain words from an analysis, you can include others. After stripping, some words you may wish to analyze may be dropped from the data set. If you wish to analyze these words anyway place them in an ASCII file in the same format as your Exclude file and enter the file name here.

A second use for the Include file is for a Zelf-analysis (named after the famous Rudolph Zelf, from Vienna). If you choose to do a Zelf-analysis the Include file contains words that may reference respondents in the text. E.G. "I", "Me", "MY", a name, or any specific identifier you used in you data collection (do not specify your Id number in this file - CATPAC will associate the Id number with the self references automatically).

File descriptor

This prompt is asking you for the filename you want **CATPAC** to use to as a prefix when labelling your output files. You may use any name you wish (up to 8 characters). When you are doing a network run **CATPAC** will automatically create several files using the filename you entered and supply an extension for each: a labels file (**.LBL**), an output file (**.CAT**), a weighted input network file (**.WIN**), and a coordinate file (**.CRD**). The labels file contains the unique words **CATPAC** identified in the analysis, while the output file contains the basic **CATPAC** outputs -- i.e., the word counts, alphabetic listing and cluster analysis. The weight input network contains the connection strengths between nodes and the coordinate file is used by **PLOT** to make a perceptual map. It contains the coordinate loadings of each unique word on a 3 dimensional set of axes. The more ambitious can actually draw their own map (by hand) using the x, y, and z coordinates listed in this file.

For example, if you were analyzing a text file that contained information about cars, you might use the filename CARS, and **CATPAC** would create the following files: **CARS.LBL** and **CARS.CAT**, **CARS.WIN**, and **CARS.CRD**. These file names are displayed automatically at the bottom of the Network analysis screen

Unique words

At this field, **CATPAC** is asking you how many words you want to carefully study. Most of the time, you will only want to use only the top fifteen, or twenty, or 30 unique words. This version of **CATPAC** can perform higher-order analyses on as many as 150 words. If you need to study more than 150 words, call TERRA; we have another version of the program that can read more words, but you will

need special instructions, and perhaps, a faster machine.

CATPAC identifies unique words in the following manner. First, the program looks at every word that occurs once or more, and then checks to see if the number of such words is greater than the number of unique words you specified. If it is, **CATPAC** will study every word that occurs twice or more; and then check to see if the number of such words is still greater than the number you requested. If it is, **CATPAC** will study those that occur three or more times, and so on, until it finally obtains the number of unique words you specified.

Many times, **CATPAC** will provide you with fewer unique words than the number requested. When this happens, it means that there were several words which occurred with the same frequency as the nth unique word and, if included, the number of unique words identified would have exceeded the number you requested. Hence, to avoid giving you more than what you ask for, **CATPAC** deletes all of these words, leaving you with fewer unique words.

A very good reason to conduct a FREQUENCY RUN (see below) prior to doing any other analysis is to examine the frequency with which the words you wish to study occur in the text. Doing this will help you determine exactly how many words you should specify at this prompt. The process is simple. Using the descending frequency list on your initial FREQUENCY RUN, examine the rank-order position of all words you want to include in your analysis. Find the rank-order position of the last word you want to include in the analysis, and specify that as the number of unique words you want **CATPAC** to study.

Remember, if there are other words which occur with the same frequency (as the last word you want to include) you must count-down to the rank-order position of the last word (with the same frequency) and specify that number as the number of unique words you wish to study.

Window size (or -1)

CATPAC works by passing a moving window of size *n* through your file. If you were to enter a window size of 7 (a good guess to start with in most cases), **CATPAC** would read your text seven words at a time. So, for example, if you were to specify a window size of 7, and a slide size of 1, CATPAC would read words 1 through 7, then words 2 through 8, then words 3 through 9, and so on.

Any time a word is in the window, the neuron representing this word becomes active. Connections among active neurons are strengthened, so words that occur close to each other in the text tend to become associated in **CATPAC**'s memory.

If you enter -1 instead, **CATPAC** abandons the moving window, and looks for -1's in your data file. These must occur in columns 1 and 2, and all the text that lies between these delimiters is considered a case.

If you use the moving window model, you do not need any -1's in your file, and **CATPAC** will make its own cases automatically using the window size you specified. Further, having -1's in your data file will not adversely effect your run if your choose to read your file using a moving window.

Slide size

This prompt is asking you how you would like the moving window to "slide" through the text. The number you select dictates how many words the window will skip prior to reading the text. You may select any increment you like. For example, if you chose a window of 5, and a slide size of 1, **CATPAC** would read words 1 through 5, 2 through 6, etc. If you chose a window of 5 and a slide of 2, **CATPAC** would read words 1 through 5, then 3 through 7, etc. Slide sizes larger than 1 are most often used when

you have a very large text file from which you want to draw "samples". This is a new field, so feel free to experiment.

If you entered -1 for window size (case by case analysis) this field is ignored.

Cycles

CATPAC's network analysis procedure works in the following manner.

When words are present in the scanning window, the neurons assigned to those words are active, and the connection among all active neurons is strengthened. But the activation of any neuron travels along the pathways or connections among neurons, and can in turn activate still other neurons whose associated words may not be in the window. These neurons can, in turn, activate still other neurons, and so on.

In an actual (biological) neural network, these processes go on in parallel and in real time, so that the signal coming into the network is spreading at different rates of speed throughout the network, and neurons are becoming active and inactive at different times. (This process of delay is called *hysteresis*.)

In a serial computer like yours, however, this is extremely difficult process to model, and so the network is updated periodically all at once. Each update is called a cycle.

Letting **CATPAC** cycle two or three times allows second and third order relationships among the words to be considered.

Very little cycling (or none at all as in the simple co-occurrence model) tends to find only very superficial associations. Too much thinking, however, is not always a good thing, since **CATPAC** can tend to see things as all pretty much alike if its allowed to cycle too many times.

> Some analysts with a warped sense of humor like to refer to this problem as "The Buddhist Monk" syndrome, since, after sufficient contemplation, it appears that all things are one.

Clamping

When a word is found in the window, its neuron is activated. But it can become de-activated again as the network goes through its normal processes, just as you (yourself) see things, become aware of them, and then forget them. (If you never forgot, your mind would become so cluttered with images in only a few minutes that you could not go on with life).

When you choose to clamp the nodes (another word for *neuron*), you prevent them from turning off again. It's like writing yourself a note and holding it in front of you so you must always pay attention to the words in the note.

Zelf-analysis

A self-analysis will allow you to identify self-references based on predetermined id's within your data set. These points will be plotted as cloud using plot. This option allows you to determine the rough boundary and location of the self-point of the text.

There are two ways to do a self-analysis. For either you have to include as the first line of each case an id number that begins with either * or +. Method I, locate by id, simply associates the Id number

of any given case with the case. Method II, locate by self-reference, will essentially replace any self-referent referred to in your Include file with the id and treat that id as a node. Since method II forces a repetition of the id, this will lead to stronger associations between the self point and the text, if the text is sufficiently "rich" in self-reference. If there are no or few self-references, the Id method ought to be used.

To do a self-analysis, press Enter on the self-analysis field. You will be given a pop-up menu of choices. Using the arrow keys, choose the method you wish and press Enter again. Remember, if you choose to do a self-analysis, your Include file becomes a self-reference file.

Network parameters

CATPAC can simulate four different kinds of neurons (functional forms), and the overall performance of **CATPAC** depends on three parameters (threshold, decay rate, and learning rate). The most generally useful neuron and some reasonable values for the three general parameters have been chosen as defaults in **CATPAC**. But you can change them if you wish, and none of these neuron types or parameters are sacred, even those selected by Terra as defaults. You might well find **CATPAC** performs better for some tasks with a different choice of neurons and/or default parameters. In order to change any defaults, just tab to the field of choice and enter a different value.

Function form

This option allows you to try different transfer functions. A true chiphead would jump at the chance to play with these. You can choose from four: a logistic varying between 0 and +1, a logistic varying between -1 and +1, a hyperbolic tangent function varying between -1 and +1, and a linear function varying between -1 and +1. Some writers speculate that different functions are better for different kinds of task, but no one knows for sure at this time.

The default threshold is 0.0. If you choose the logistic function that varies between 0 and 1, the threshold will automatically be set to .5. If you'd like to experiment with different transfer functions, press enter at this field and you will get a menu of the four forms, arrow to the function of your choice and press enter.

> Note: A Chiphead is a person with an exceptional commitment to computing. If you plan to do basic research on various transfer functions, you are one.

Threshold

Each neuron in **CATPAC** is either turned on by being in the moving window, or else receives inputs from other neurons to which it is connected. These inputs are transformed by a *transfer function*.

After the inputs to any neuron have been transformed by the transfer function, they are summed, and, if they exceed a given threshold, that neuron is activated; otherwise it remains inactive.

The default threshold for the three transfer functions that vary between ± 1 is 0.0, and .5 for the logistic varying between 0 and +1. By lowering the threshold, you make it more likely for neurons to become activated; by raising the threshold, you make it less likely for neurons to become activated.

Decay rate

When you see an object, neurons which represent that object are activated. When the object is gone,

the neurons (fortunately) turn off again. (If they didn't, you'd be seeing everything you ever saw all the time.) The decay rate specifies how quickly the neurons return to their rest condition (0.0) after being activated. The default rate is .9, which means that each neuron, if not reactivated, will lose 90% of its activation each cycle. Raising the rate makes them turn off faster; lowering the rate means they are likely to stay on longer.

Learning rate

When neurons behave similarly, the strength of the connection between them is strengthened. The learning rate is how much they are strengthened in each cycle. Default is .001. Increasing this rate makes **CATPAC** learn faster. Faster is not always better, though, since too high of a rate can make **CATPAC** oscillate back and forth as new information is read. No one knows the optimum rate, or even if there is an optimum rate, so feel free to experiment.

Output files

As explained above **CATPAC** automatically produces its out files for a Network analysis. These four fields: **Catpac file, Label file, Win file, and Coordinate file** merely display the names of these files based on the File descriptor you entered. These fields can only be altered by changing the contents of the File descriptor field.

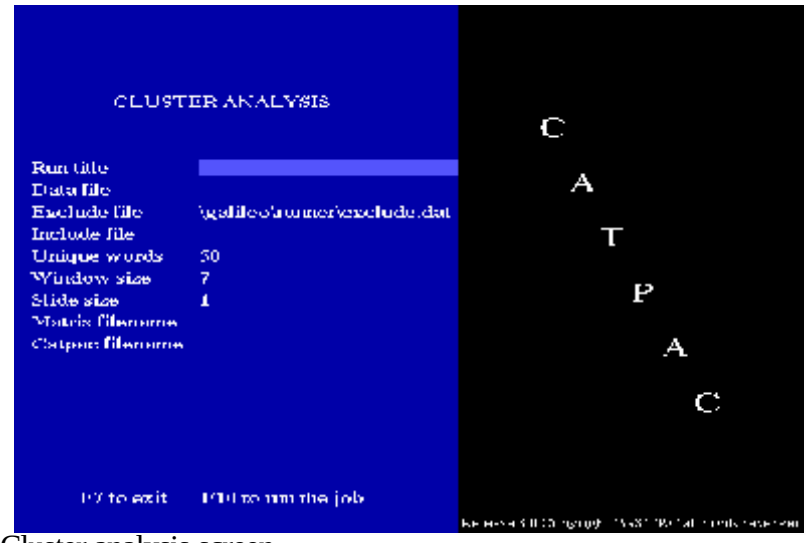
COMPLETED SCREEN

A completed Network analysis screen might look like this:

4 Completed Network screen

CLUSTER ANALYSIS

To do a Cluster analysis, press **F2** at the Main Menu. The cluster analysis will appear:



5 Cluster analysis screen

Many of the fields for a Cluster analysis are identical to those for a Network analysis. The main differences are a lack of network parameters and how you enter the output file names.

Matrix file

Cursor down to the Matrix file field and enter the name of a file to store your co-occurrences. If you enter a file name at this field, **CATPAC** will output a file that contains a list of word co-occurrences it encountered within the window-size you specified. CATPAC will list every co-occurrence, and tell you how many times it encountered each co-occurrence. If you leave this field blank **CATPAC** will not produce this output.

Catpac file

You must enter a file name for this field. This is the same output mentioned above (**.CAT**) and is the standard output for **CATPAC**. You may call it anything you like, but we suggest you use the **.CAT** extension. If you leave this field blank, **CATPAC** will not run and will prompt you for a file name.

FREQUENCY ANALYSIS

CATPAC has the capability to read every word in your file and list these words in descending order of frequency, as well as in alphabetical order. A FREQUENCY RUN is typically used by analysts to help them "clean" their data prior to performing advanced analyses. A FREQUENCY RUN can help you find typographical errors, synonyms, plurals, pro-nouns, and other such words that you may want to recode using a word processing program (or a text editor) prior to proceeding with further analyses.

If you want to perform a Frequency analysis, type **F1** at the main menu. Here is the screen you'll see:

This is a very simple screen. Most of the information requested is explained above. You must supply an output file name for your frequencies. Cursor to the **Output file** field and enter a file name to store this output on.

6 Frequency analysis screen

HELP

You may get help at any time by pressing **F3**. There are two ways to get help. If you press F3 from the main menu, you will be given the option to press any other function key for help on any of CATPAC's main operations. When you are done with help on the main menu, press F7 as the screen indicates.

If you are filling out any specific analysis screen, press F3 on any field and you will get a brief explanation of what the field is and what kind of information it is looking for.

RESET TO DEFAULTS

If you have altered any of CATPAC's default settings and wish to return to these values, press **F5** and CATPAC will reset all changed parameters to the ones supplied with the program.

INPUT TO CATPAC

CATPAC can read any text file that has been converted to ASCII. Some examples of text files people have studied using **CATPAC** include: answers to open-end survey questions, focus group transcripts, newspaper and magazine articles down-loaded from a data base, comments left on a customer telephone hot-line, and restaurant/hotel/airline comment cards.

When preparing data for a **CATPAC** analysis, keep in mind that while **CATPAC** is quite an amazing technology, it is still quite primitive. In as much, if you are studying focus group transcripts for example, you should probably first parse the file into discrete topic-specific sections, rather than have **CATPAC** try to study a file that spans 10-15 topics.

Figure 1 shows a text derived from some interviews where people were asked to describe the difference between a select set of pizza restaurants. Asking people to describe the difference between products is usually a good method, since they then usually report attributes which make a difference, instead of attributes which all the products might share.

The reader will note that this particular text file is not very long or of very high quality. Hopefully, your data set will be a little better!

```
   I like pizza, hot and fresh. I like quick delivery, like Domino's gives, but
I need quality like pizzahut. Little Caesar's is inexpensive, but I guess
pizzahut has quality. Domino's delivers, but Domino's is expensive. Little
Caesar's is inexpensive, and you get two at Little Caesar's. Little Caesar's two
for one deal is inexpensive. I like good flavor, like pizzahut, but I guess
Domino's is faster. Sometimes you want it faster, and Domino's is faster. If you
want good flavor, Pizzahut is for you, but if you want it inexpensive, Little
Caesar's is the best. It's good, Little Caesar's is good, but Pizza Hut is good
too. Domino's is not as good, but fast. Domino's is fast.  I think Domino's has
fast delivery, and Domino's fast delivery means a lot to me. Pizzahut's quality
is important, but it's not worth it; Little Caesar's two for one is really good.
Two for one? Little Caesar's is the two for one place. Pizzahut quality sets it
apart, but Little Caesar's is inexpensive. Pizzahut is expensive. But of course
Domino's fast delivery can be important. When you want fast delivery, Domino's
is the fast delivery place.     For inexpensive pizza, Little Caesar's is most
inexpensive of all. Inexpensive little caesar's is the place for two for one:
little caesar's two for one. Little Caesar's is inexpensive.
```

Figure 1 PIZZA INTERVIEWS

CATPAC OUTPUTS

Getting **CATPAC** to analyze these interviews is very simple. In this case, we asked **CATPAC** to cycle once, and to identify no more than 20 unique words. We set the window size to 5, no other values were re-set. The results are shown in Figures 2 and 3.

```
CATPAC_PC  v3.00

                          05/25/93      09:29:40

TITLE:     Pizza interviews
DATA FILE: PIZZA.TXT

TOTAL WORDS           115     THRESHOLD            .000
TOTAL UNIQUE WORDS     17     RESTORING FORCE      .100
TOTAL WINDOWS         138     CYCLES                 1
TOTAL LINES            21     FUNCTION              Sigmoid (-1 - +1)
WINDOW SIZE             5     CLAMPING              Yes
SLIDE SIZE              1
```

DESCENDING FREQUENCY LIST

WORD	FREQ	PCNT	CASE FREQ	CASE PCNT
LITTLE	13	11.3	58	42.0
CAESAR	13	11.3	57	41.3
DOMINO	11	9.6	46	33.3
INEXPENSIVE	9	7.8	37	26.8
PIZZAHUT	7	6.1	35	25.4
TWO	7	6.1	32	23.2
GOOD	7	6.1	28	20.3
FAST	7	6.1	26	18.8
LIKE	6	5.2	21	15.2
DELIVERY	6	5.2	26	18.8
YOU	6	5.2	26	18.8
ONE	6	5.2	27	19.6
QUALITY	4	3.5	20	14.5
WANT	4	3.5	20	14.5
PIZZA	3	2.6	12	8.7
FASTER	3	2.6	11	8.0
PLACE	3	2.6	15	10.9

ALPHABETICALLY SORTED LIST

WORD	FREQ	PCNT	CASE FREQ	CASE PCNT
CAESAR	13	11.3	57	41.3
DELIVERY	6	5.2	26	18.8
DOMINO	11	9.6	46	33.3
FAST	7	6.1	26	18.8
FASTER	3	2.6	11	8.0
GOOD	7	6.1	28	20.3
INEXPENSIVE	9	7.8	37	26.8
LIKE	6	5.2	21	15.2
LITTLE	13	11.3	58	42.0
ONE	6	5.2	27	19.6
PIZZA	3	2.6	12	8.7
PIZZAHUT	7	6.1	35	25.4
PLACE	3	2.6	15	10.9
QUALITY	4	3.5	20	14.5
TWO	7	6.1	32	23.2
WANT	4	3.5	20	14.5
YOU	6	5.2	26	18.8

Figure 2 CATPAC WORD COUNTS

Run Summary

Figure 2 shows the most basic output of **CATPAC**. It consists of a summary of the parameters selected, and a frequency count of the main words found in the text. It shows that there were 115 total words in the text, and that 17 unique words were found. There were 138 windows in the analysis, and 21 lines of text.

The left-most columns present the major words in descending order of frequency of occurrence. They show that "Little" was the most frequently occurring word, that it occurred 13 times, which was 11.3% of all occurrences. "Little" appeared in 58 or 42.0% of the scanned windows. This last figure is referred to in the output as a "CASE FREQ" and indicates the number of times a given word appears in a case. If you had delimited each response with a -1 and done a case-by-case analysis, The "CASE FREQ" would indicate the number of respondents to mention each word. The words "pizza" "faster" and "place" occurred least often, three times each. **CATPAC** didn't consider any words that occurred fewer than three times, since that would have resulted in the identification of more than the 20 unique words we requested.

The right-most columns give exactly the same information as the left-most columns, except the unique words are now listed in alphabetical order for easy look-up.

Dendogram

Figure 3 shows the output from the hierarchical cluster analysis. These pictures are called "dendograms,", and they look a bit like the skyline of a city seen from afar. The "buildings" underneath the words show which words cluster together.

As can be seen in Figure 3, this cluster analysis reflects the information contained in the text quite well. The words "little" and "caesar" cluster very sharply together as we would expect. "Domino" and "Fast" cluster very closely, and the word "delivery" joins this cluster at a slightly lower level (Domino's specializes in fast delivery). Similarly, "Pizzahut", "quality", and like form a third cluster.

Figure 3 CLUSTER ANALYSIS DENDOGRAM

As we move downward through the diagram, each of the clusters grows larger, including more and more terms. Eventually, the first cluster includes "Domino," "fast," "delivery," "you," along with the subcluster "you", "want", and "faster." Little Caesar ends up in a cluster which includes "little" "caesar","inexpensive," and "place" with the subcluster "two" and "one" (Little Ceaser offers two for one pizzas).

Weight input networks (.WIN file)

As we've said, **CATPAC** works by finding the connections among a set of neurons which stand for the main words in the text. This network of interconnected neurons is the main product of **CATPAC**'s thinking, and it is stored in a file with the suffix *.WIN*. This matrix is not of much use by itself, but it contains all the information **CATPAC** has learned about the text, and can serve as an input to other programs, including **ORESME**™. **ORESME** simulates the processes a neural network carries out when it receives inputs of various kinds of information, and can give an interesting simulation of how individuals or groups of people might respond to different messages.

Since the .WIN matrix is in a standard ASCII format with its format statement in front of the data, it may be input into a variety of standard statistical and mathematical analyses software packages.

Coordinates (the .CRD file)

When you request a network analysis, CATPAC will also produce a file of spatial coordinates which has the same generic name as the other files produced by CATPAC, but which ends in the extension *.CRD*. This file contains information which can be used to generate a pictorial representation of the word associations CATPAC discovered during its analysis of the text.

The .CRD file contains the coordinates of the words on the basis of which plots can be made. To make a perceptual map the user must call up the program PLOT (which is also provided with your software) and type this file name with the .CRD extension after the F1 prompt. This .CRD file provides the basis for a wide variety of analysis of **CATPAC** data, including perceptual maps, development of marketing and advertising strategies, and tracking of perceptual change.

Figure 4 shows how the same clusters portrayed in the dendogram above, this time in the form of a perceptual map. To make the plot more readable, we used an option in PLOT that allows the user to remove concepts from the display. In this case we retain only the words in the major clusters.

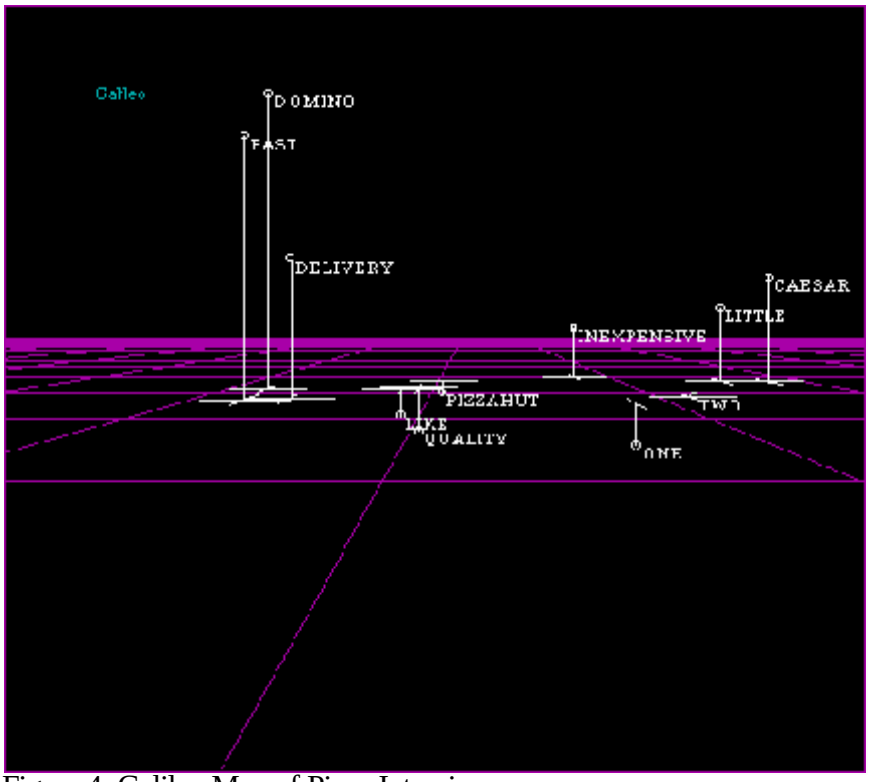


Figure 4 Galileo Map of Pizza Interviews

Notice at the top of the plot the three concepts DOMINO'S, FAST, and DELIVERY. To the right and lower, LITTLE, CEASER, INEXPENSIVE, ONE and TWO. Right in the center are the terms PIZZAHUT, QUALITY, and LIKE.

This perceptual map provides not only an alternative way to represent the same results as the dendogram, but it allows for a wide range of special analyses.

Figure 5 shows the same perceptual map in 2 dimensions.

For a detailed description of these features, refer to the Galileo*PLOT manual provided with your diskette.

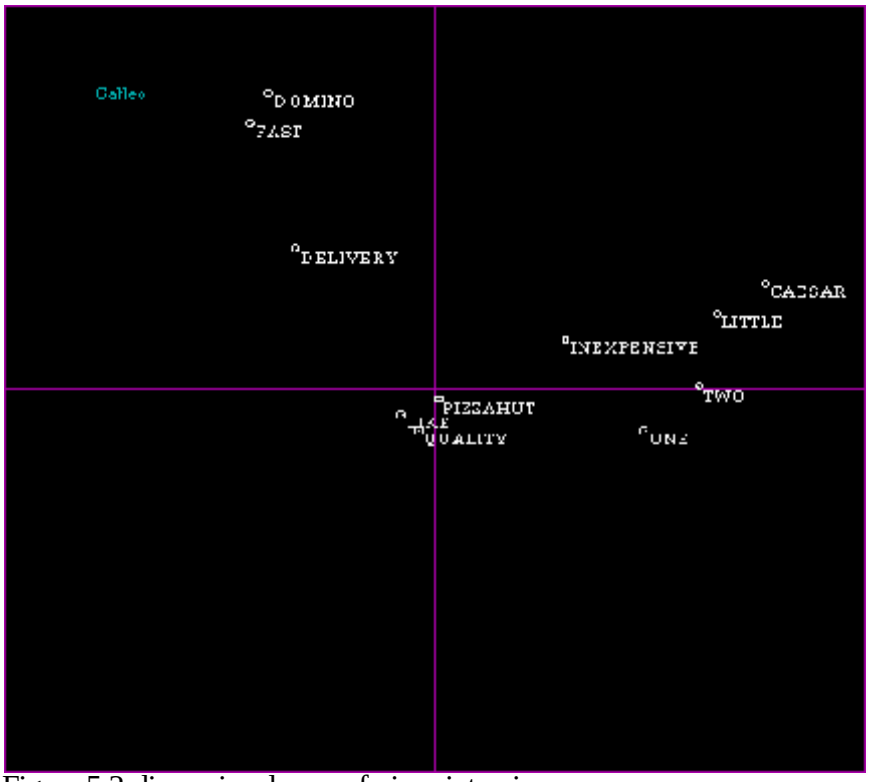


Figure 5 2-dimensional map of pizza interviews

SOME FINAL REMARKS

CATPAC represents a new generation of artificial neural software that can do things older computer software couldn't. In this manual we've tried to acquaint you with some of the new possibilities this technology makes available. But neural technology is so new that not even the development community has a good understanding of what's possible yet. Your best strategy is to spend time with the program and experiment. If you have any problems, please call your Terra representative.

Appendix 1: Tools

Your Galileo installation includes a directory called GALILEO\TOOLS. On this directory Terra has supplied three helpful DOS tools. First is a simple read only editor called **LOOK**. **LOOK** is a public-domain program which allows you to examine the contents of any file interactively. It is convenient since you can page up and down or scroll up, down, left and right in the file using the cursor control keys. You can also easily read the 132 column format files that V55 writes. And, since **LOOK** is a read only editor, you don't run the risk of altering important files.

To use **LOOK**, simply enter the command

LOOK [filename]

at the DOS prompt. To leave **LOOK**, press [ESC].

Also included is a very powerful ASCII editor, **EDWIN**. **EDWIN** is a public domain program which follows the formats of WORDSTAR, and can be very helpful in modifying files produced by V55 for use in the other Galileo programs and vice versa. **EDWIN** has complete online help, accessed by pressing **F2** once in the program. To start EDWIN, simply enter the command

EDWIN

at the DOS prompt. You can also enter a file directly with EDWIN by entering the command

EDWIN [filename] .

If you already have an ASCII editor you favor, you may use that instead of **EDWIN**. For more information on installing and using EDWIN, consult the documentation provided on the \GALILEO\TOOLS directory.

The last tool provided is called **UP**. Up lets you climb up your directory tree in only three keystrokes. If your default directory, for example, is GALILEO\DATA, then issuing the command

UP

at the DOS prompt will set your default directory to \GALILEO. Issuing the command again will move you to the root directory.

All three of these utilities are public domain software and are neither warranted nor supported by The Galileo Company, The Galileo Company and Computing or any of their agents. The are provided at no charge as a convenience for the user.

Note that authors of public domain software sometimes request voluntary payments from users for the use of their programs. No such payments have been made on your behalf by Terra, Galileo or any of their representatives.