User's Guide

Catpac II™

Version 2.0

Joseph K. Woelfel

Rah Press

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 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.

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 GALILEO.

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 THE GALILEO COMPANY.

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

YOU MAY MAKE REASONABLE BACKUP OR ARCHIVAL COPIES OF 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 Galileo is unable to correct it within 90 days of receipt of your report of the defect, you may return the software and Galileo 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 AND/OR INABILITY TO USE THE PROGRAM, EVEN IF THE GALILEO COMPANY OR AN AUTHORIZED GALILEO 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 of 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, 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.

CONTENTS

Chapter 1 Introduction

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 patterns of similarity based on the way they're used in text. It 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 are the many clustering algorithms provided by **Catpac**, as well as perceptual mapping provided by **ThoughtView**, and interactive clustering provided by **Oresme**.

Installing Catpac

- Place the diskette in the floppy drive.

- At the Run prompt, type A:CATPAC2 where A is the letter of the floppy drive.
- To install ThoughtView, insert the ThoughtView disk into the drive and type A:SETUP.

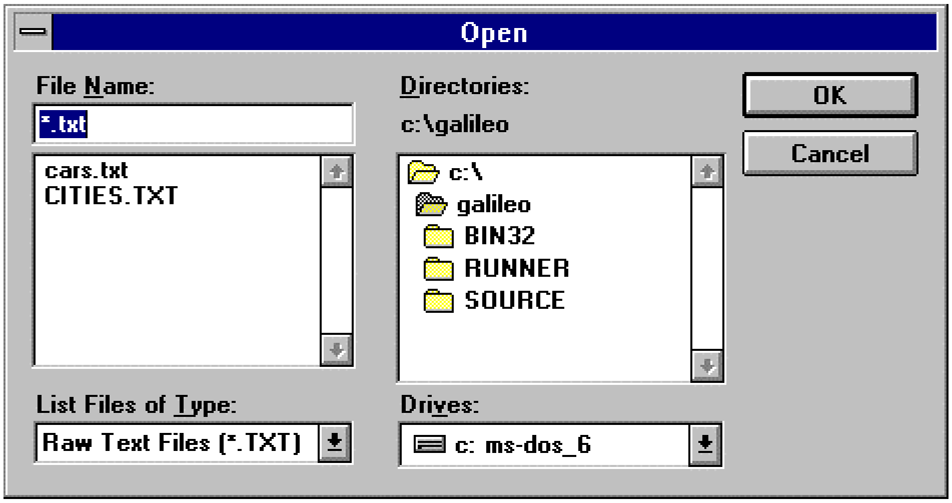
That’s it! Except a few questions that you have to answer, the install program will take care of everything.

Chapter 2 Using Catpac

Making a Dendogram

Let's start with the simplest analysis you can do.   First, press the input file button. It looks like this:

This will bring up an open file dialog box similar to the following:



File Name: *.txt

cars.txt
CITIES.TXT

Directories: c:\galileo

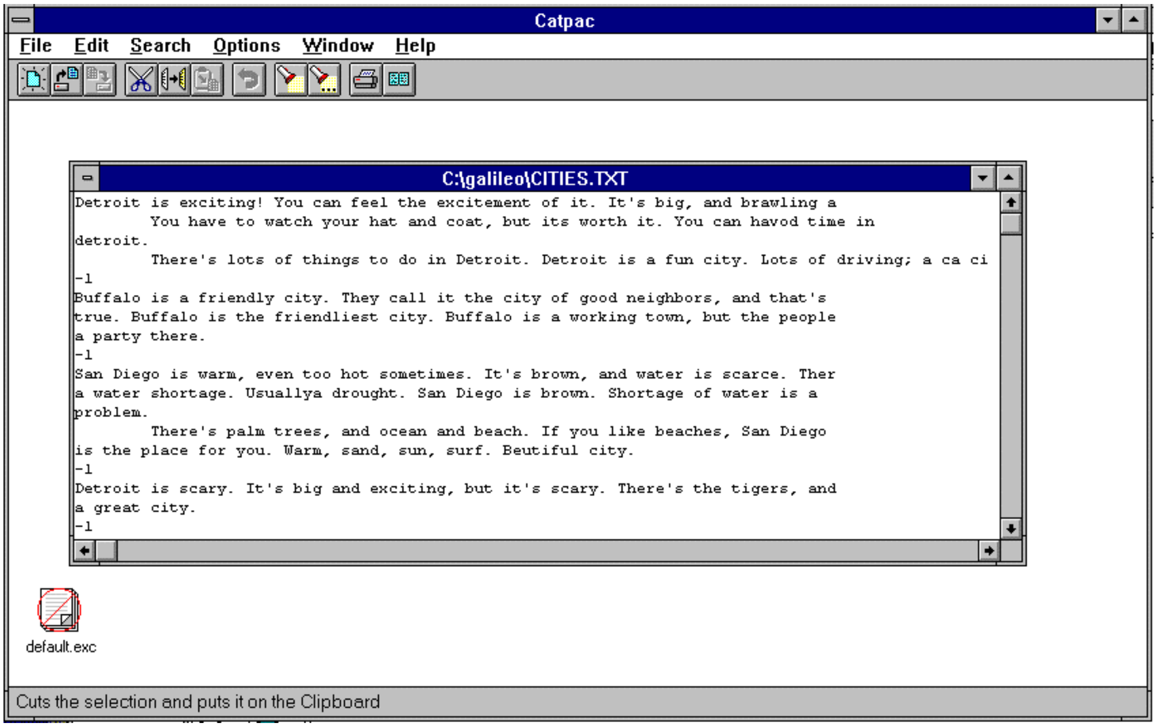
c:\
galileo
BIN32
RUNNER
SOURCE

List Files of Type: Raw Text Files (*.TXT)

Drives: c: ms-dos_6

Select the file from the file list.   By default, only files with the extension ***.txt*** are shown; this is because **Catpac** only reads input files in ASCII text form[1]. You can change the directory by double-clicking on the folders in the directory portion of the dialog box. If you select CITIES.TXT from the list your next screen will look similar to the following:

[1] Sometimes, people make the error of exiting their word processor without converting the file to ASCII. Don't make this mistake!



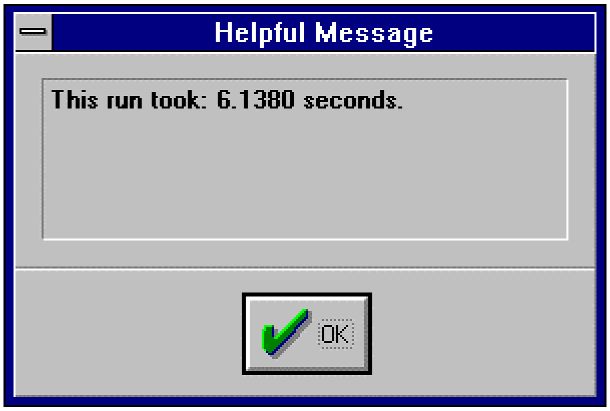
```
Detroit is exciting! You can feel the excitement of it. It's big, and brawling a
        You have to watch your hat and coat, but its worth it. You can havod time in
detroit.
        There's lots of things to do in Detroit. Detroit is a fun city. Lots of driving; a ca ci
-1
Buffalo is a friendly city. They call it the city of good neighbors, and that's
true. Buffalo is the friendliest city. Buffalo is a working town, but the people
a party there.
-1
San Diego is warm, even too hot sometimes. It's brown, and water is scarce. Ther
a water shortage. Usuallya drought. San Diego is brown. Shortage of water is a
problem.
        There's palm trees, and ocean and beach. If you like beaches, San Diego
is the place for you. Warm, sand, sun, surf. Beutiful city.
-1
Detroit is scary. It's big and exciting, but it's scary. There's the tigers, and
a great city.
-1
```

Cuts the selection and puts it on the Clipboard

To do the simplest analysis, all you need to do is select **Run|Make Dendogram** from the **File** menu. However, most users find it easier to simply place the mouse on top of the input file window and click once on the right mouse button. This will produce a pop-up menu with only one item: **Make Dendogram**. Selecting it will also run the analysis[2].

A pacification box will temporarily appear, letting you know of **Catpac**'s progress. You will then see a message similar to the following, telling you how long the run took:

[2] You can always place the mouse over any open window and press the right mouse button. It will always produce a pop-up menu with appropriate analysis choices for that particular window.



This run took: 6.1380 seconds.

After you press the **OK** button, a third window will pop up; it will contain textual information similar to the following:

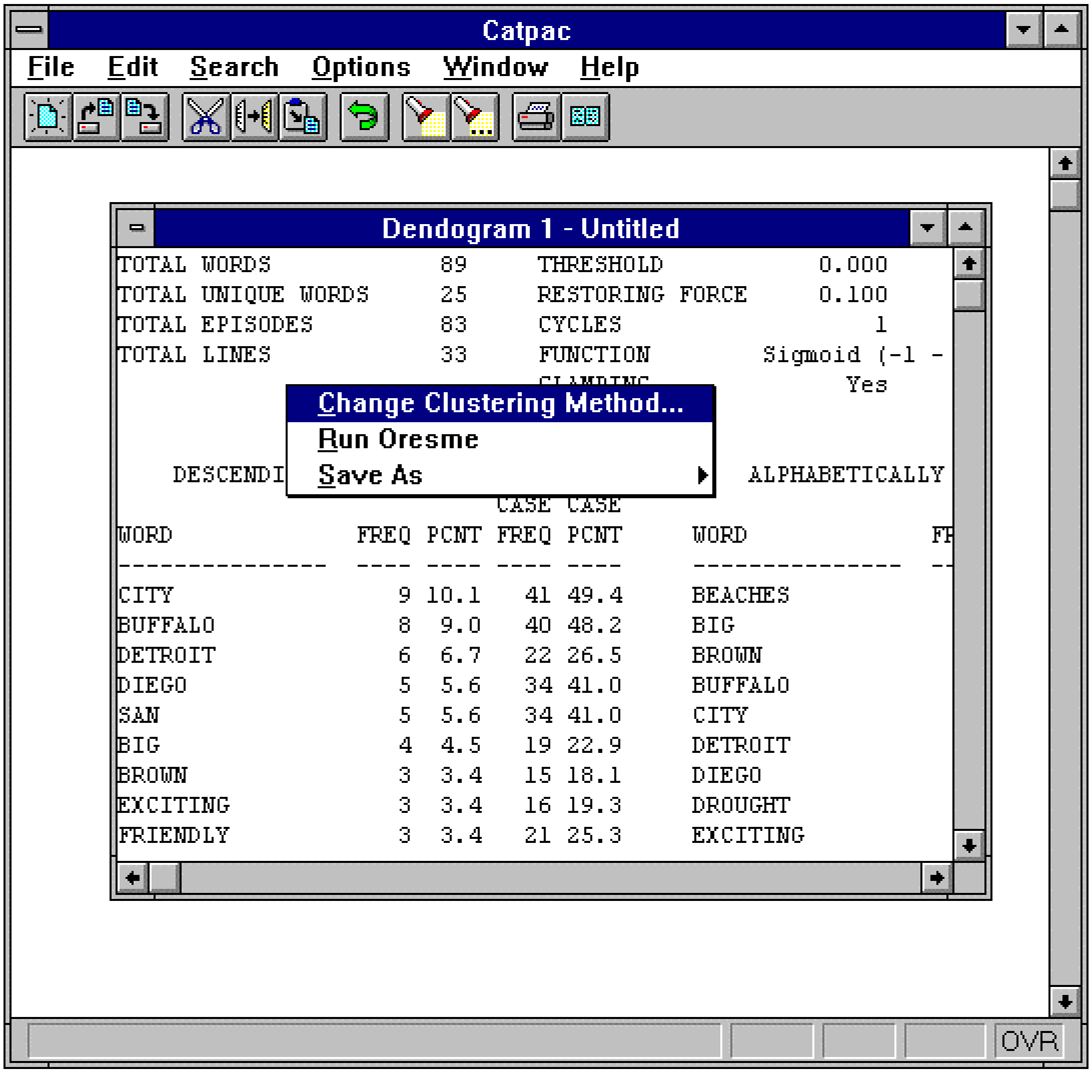
```
TOTAL WORDS            89     THRESHOLD           0.000
TOTAL UNIQUE WORDS     25     RESTORING FORCE     0.100
TOTAL EPISODES         83     CYCLES                  1
TOTAL LINES            33     FUNCTION        Sigmoid (-1 - +1)
                              CLAMPING              Yes

    DESCENDING FREQUENCY LIST                ALPHABETICALLY SORTED LIST
                           CASE CASE                                    CASE CASE
WORD             FREQ PCNT FREQ PCNT     WORD             FREQ PCNT FREQ PCNT
---------------  ---- ---- ---- ----     ---------------  ---- ---- ---- ----
CITY                9 10.1   41 49.4     BEACHES             2  2.2   14 16.9
BUFFALO             8  9.0   40 48.2     BIG                 4  4.5   19 22.9
DETROIT             6  6.7   22 26.5     BROWN               3  3.4   15 18.1
DIEGO               5  5.6   34 41.0     BUFFALO             8  9.0   40 48.2
SAN                 5  5.6   34 41.0     CITY                9 10.1   41 49.4
BIG                 4  4.5   19 22.9     DETROIT             6  6.7   22 26.5
BROWN               3  3.4   15 18.1     DIEGO               5  5.6   34 41.0
EXCITING            3  3.4   16 19.3     DROUGHT             2  2.2    8  9.6
FRIENDLY            3  3.4   21 25.3     EXCITING            3  3.4   16 19.3
FUN                 3  3.4   15 18.1     FALLS               2  2.2   14 16.9
GOOD                3  3.4   21 25.3     FRIENDLY            3  3.4   21 25.3
NEIGHBORS           3  3.4   18 21.7     FUN                 3  3.4   15 18.1
NIAGARA             3  3.4   15 18.1     GOOD                3  3.4   21 25.3
PALM                3  3.4   18 21.7     LOTS                2  2.2   11 13.3
SAND                3  3.4   20 24.1     NEIGHBORS           3  3.4   18 21.7
SUN                 3  3.4   21 25.3     NIAGARA             3  3.4   15 18.1
SURF                3  3.4   19 22.9     PALM                3  3.4   18 21.7
THERE'S             3  3.4   19 22.9     PARTY               2  2.2   14 16.9
TREES               3  3.4   17 20.5     SAN                 5  5.6   34 41.0
WATER               3  3.4   13 15.7     SAND                3  3.4   20 24.1
BEACHES             2  2.2   14 16.9     SUN                 3  3.4   21 25.3
DROUGHT             2  2.2    8  9.6     SURF                3  3.4   19 22.9
FALLS               2  2.2   14 16.9     THERE'S             3  3.4   19 22.9
LOTS                2  2.2   11 13.3     TREES               3  3.4   17 20.5
PARTY               2  2.2   14 16.9     WATER               3  3.4   13 15.7
```

Following these basic statistics is a dendogram that describes the relationships between the most commonly occurring concepts:

```
B P T S S S B D S W D B D E L T F B C F G N F N P
E A R A U U R I A A R I E X O H U U I R O E A I A
A L E N N R O E N T O G T C T E N F T I O I L A R
C M E D . F W G . E U . R I S R . F Y E D G L G T
H . S . . . N O . R G . O T . E . A . N . H S A Y
E . . . . . . . . . H . I I . ' . L . D . B . R .
S . . . . . . . . . T . T N . S . O . L . O . A .
. . . . . . . . . . . . . G . . . . . Y . R . . .
. . . . . . . . . . . . . . . . . . . . . S . . .
. . . . . . . . . . . . . . . . . . . . . . . . .
. . . . . . . . . . . . . . . . . . . . . . . . .
. . . . . . . . . . . . . . . . . ^^^ . . . . . .
. . . . . . . . . . . ^^^ . . . . ^^^ . . . . . .
. . . . . . . . . . . ^^^^^ . . . ^^^ . . . . . .
. . . . . . . . . . . ^^^^^^^ . . ^^^ . . . . . .
. . . . . . . . . . . ^^^^^^^^^ . ^^^ . . . . . .
. . . . . . . ^^^ . . ^^^^^^^^^ . ^^^ . . . . . .
. . . . . . . ^^^ . . ^^^^^^^^^ . ^^^^^ . . . . .
. . . . . . . ^^^^^ . ^^^^^^^^^ . ^^^^^ . . . . .
. . . . . . . ^^^^^ . ^^^^^^^^^ . ^^^^^ ^^^ . . .
. . . . . . . ^^^^^ . ^^^^^^^^^ . ^^^^^^^^^ . . .
. . . ^^^ . . ^^^^^ . ^^^^^^^^^ . ^^^^^^^^^ . . .
. . . ^^^ . . ^^^^^ . ^^^^^^^^^ . ^^^^^^^^^ ^^^ .
. . . ^^^ . ^^^^^^^ . ^^^^^^^^^ . ^^^^^^^^^ ^^^ .
. . . ^^^^^ ^^^^^^^ . ^^^^^^^^^ . ^^^^^^^^^ ^^^ .
. . . ^^^^^ ^^^^^^^ . ^^^^^^^^^^^ ^^^^^^^^^ ^^^ .
. ^^^ ^^^^^ ^^^^^^^ . ^^^^^^^^^^^ ^^^^^^^^^ ^^^ .
^^^^^ ^^^^^ ^^^^^^^ . ^^^^^^^^^^^ ^^^^^^^^^ ^^^ .
^^^^^ ^^^^^ ^^^^^^^ . ^^^^^^^^^^^ ^^^^^^^^^^^^^ .
^^^^^ ^^^^^ ^^^^^^^^^ ^^^^^^^^^^^ ^^^^^^^^^^^^^ .
^^^^^^^^^^^ ^^^^^^^^^ ^^^^^^^^^^^ ^^^^^^^^^^^^^ .
^^^^^^^^^^^ ^^^^^^^^^ ^^^^^^^^^^^ ^^^^^^^^^^^^^^^
^^^^^^^^^^^^^^^^^^^^^ ^^^^^^^^^^^ ^^^^^^^^^^^^^^^
```

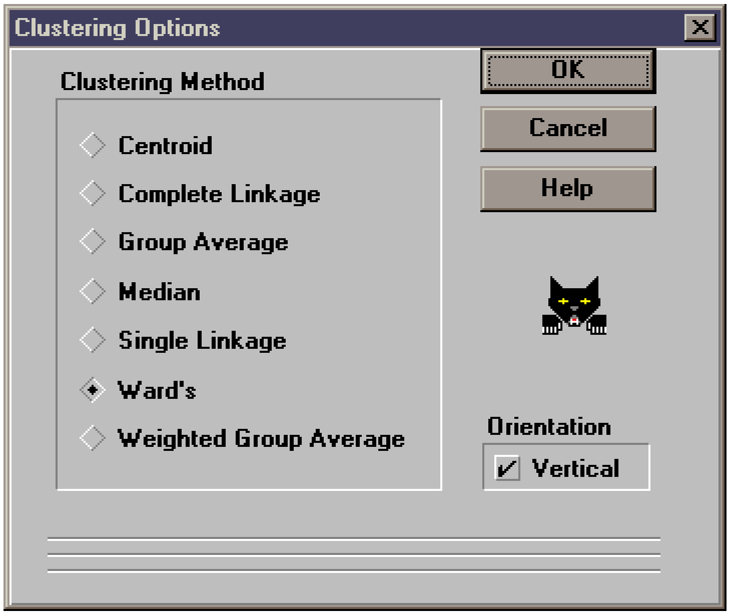
Changing the Clustering Method.

The above dendogram was constructed using the “group average method.” Six other common clustering methods are also available. You can change these on the fly by placing the mouse pointer above the dendogram window, clicking on the right mouse button, and selecting **Change Clustering Method** from the following speed menu:



```
TOTAL WORDS             89      THRESHOLD          0.000
TOTAL UNIQUE WORDS      25      RESTORING FORCE    0.100
TOTAL EPISODES          83      CYCLES                 1
TOTAL LINES             33      FUNCTION      Sigmoid (-1 -
                                CLAMPING             Yes


    DESCENDI                                 ALPHABETICALLY
                              CASE  CASE
WORD              FREQ  PCNT  FREQ  PCNT    WORD            FR
---------------   ----  ----  ----  ----    ---------------  --
CITY                 9  10.1    41  49.4    BEACHES
BUFFALO              8   9.0    40  48.2    BIG
DETROIT              6   6.7    22  26.5    BROWN
DIEGO                5   5.6    34  41.0    BUFFALO
SAN                  5   5.6    34  41.0    CITY
BIG                  4   4.5    19  22.9    DETROIT
BROWN                3   3.4    15  18.1    DIEGO
EXCITING             3   3.4    16  19.3    DROUGHT
FRIENDLY             3   3.4    21  25.3    EXCITING
```

This brings up the following dialog box:



Clustering Method

◇ Centroid
◇ Complete Linkage
◇ Group Average
◇ Median
◇ Single Linkage
◆ Ward's
◇ Weighted Group Average

Orientation

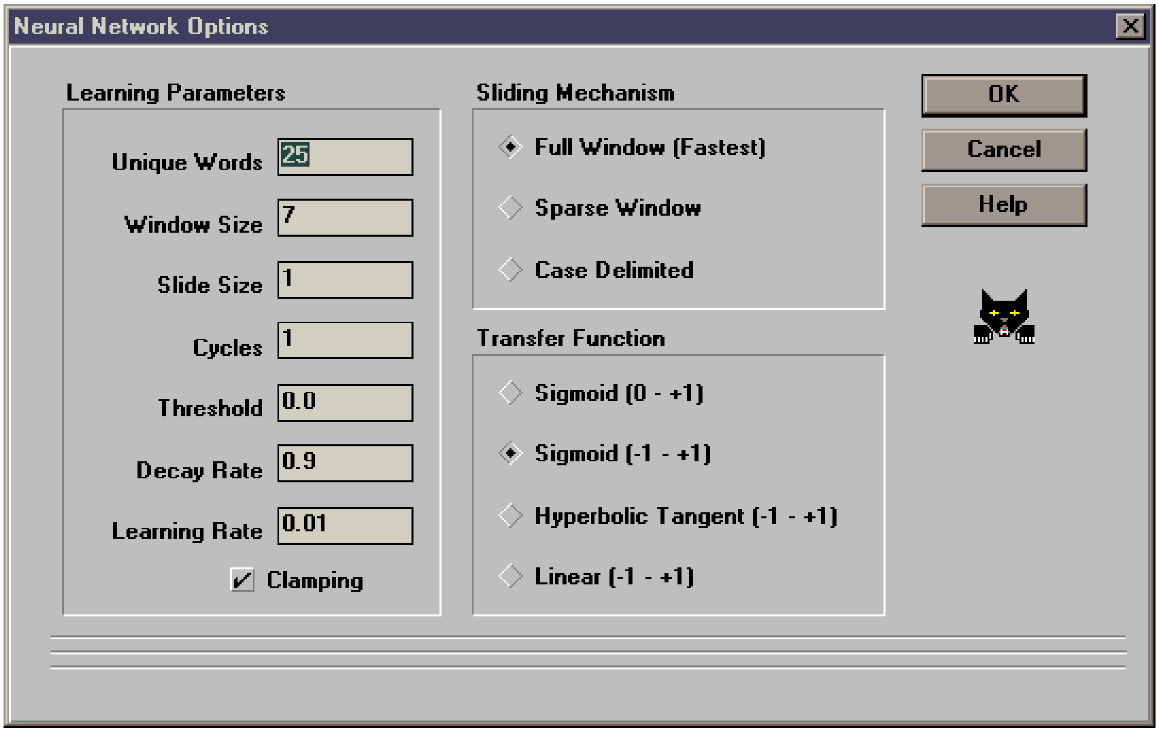
☑ Vertical

Select the clustering method of your choice and press **OK**. **Catpac** will quickly re-cluster your dendogram window without having to go through the more time-consuming training process. If you would like to know more about the advantages and disadvantages of the different clustering algorithms, select **Help**.

You can get the same dialog box by selecting **Options|Clustering Method**, but this only changes the default, and will not affect any prior analyses.

Network Options

To change the network options, select **Options|Neural Network**. This will call up the following dialog box:



Neural Network Options

Learning Parameters

Unique Words 25

Window Size 7

Slide Size 1

Cycles 1

Threshold 0.0

Decay Rate 0.9

Learning Rate 0.01

☑ Clamping

Sliding Mechanism

● Full Window (Fastest)

○ Sparse Window

○ Case Delimited

Transfer Function

○ Sigmoid (0 - +1)

● Sigmoid (-1 - +1)

○ Hyperbolic Tangent (-1 - +1)

○ Linear (-1 - +1)

Any changes made here will affect all subsequent analyses for this session.

Unique words

This is the number of unique words that you want in your analysis. It's also the number of words that you will see in your dendogram. The network will use fewer words if there aren't enough of them in the data, and more if you add some using an include file. **Catpac** selects words based on frequency. The most frequent words are chosen first.

Most of the time, you will only want to use the top 15 to 30 unique words. This version of **Catpac** can perform higher-order analyses on as many as 160 words. If you need to study more than 160 words, call Galileo; we have a Windows NT version of **Catpac** that can handle any number of words, and can do them twice as fast.

Window size

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 7 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.

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.

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. In addition, 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 an extremely difficult process to model, and so the network is updated periodically all at once. Each update is called a cycle.

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 it's 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.

Catpac’s default value is 1 cycle, and most analysts find that works well in most cases. But don’t be afraid to experiment.

Clamping

When a word is found in the window, its neuron is activated. However, 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.

Chip-head [3] network options

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; none of these neuron types or parameters are sacred, even those selected by Galileo 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 chip-head 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 tasks, but no one knows for sure at this time.

[3] A Chip-head 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 is zero. 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.

Changing the threshold is perhaps most useful when running Oresme, since it can change the number of responses Oresme makes. Raising the threshold decreases the number of associations Oresme finds.

Decay rate

When you see an object, neurons that 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.

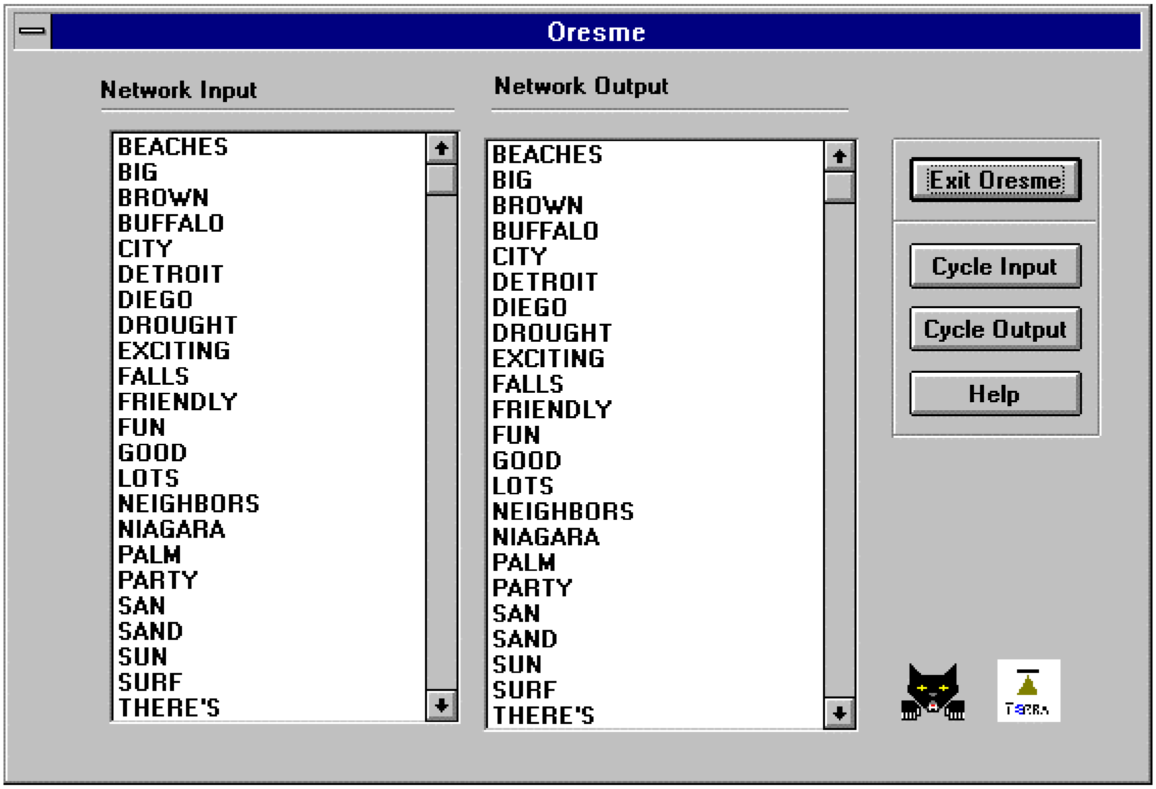
Learning rate

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. The default is .01. Increasing this rate makes Catpac learn faster. Faster learning rates are not always better, though, since too high a rate can make Catpac oscillate back and forth as new information is read, and can lead to the “Buddhist Monk Effect.”

The default rate works well with most files. If your files are lengthy, you might want to lower the default learning rate perhaps to .001. (If you have one very large cluster and one small one; if the right edge of the first cluster tails off in a straight line, or if your dendogram looks like a mitten instead of a glove, the chances are you need to either raise or lower the learning rate by perhaps a factor of 10.) No one knows the optimum rate, or even if there is an optimum rate, so feel free to experiment

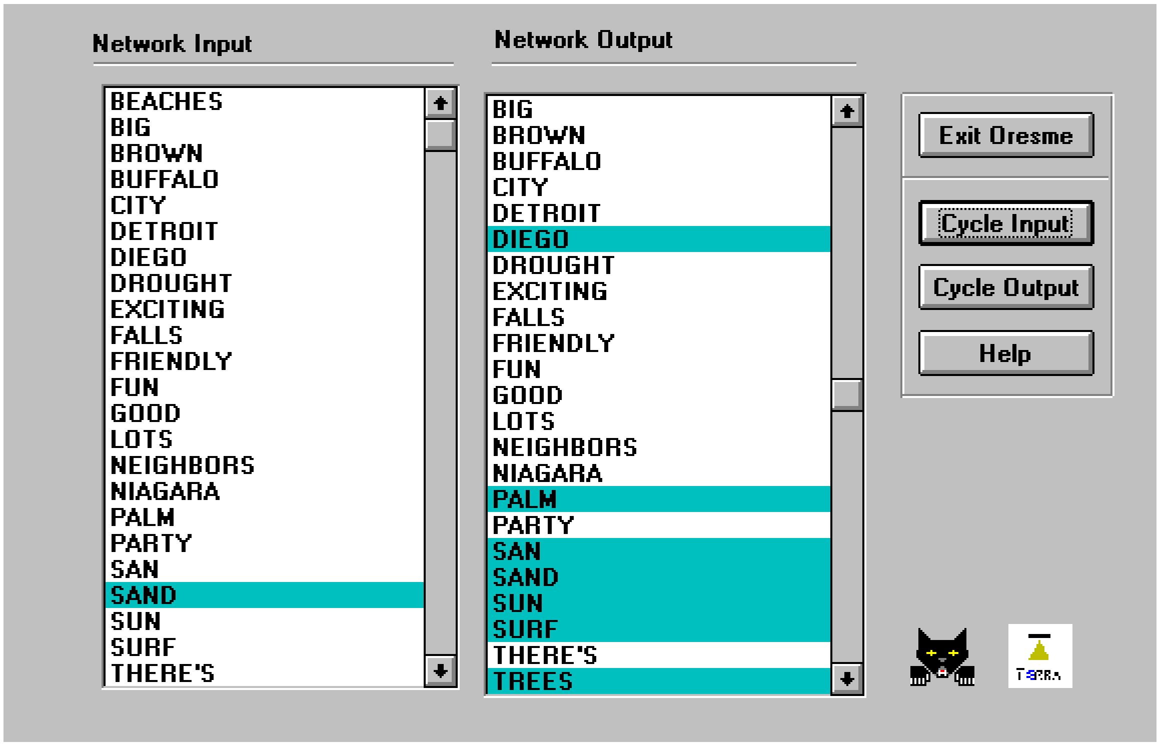
Chapter 3 Oresme Interactive Clustering

To run Oresme position the mouse pointer over the dendogram window of interest and click the right mouse button once. This produces a menu with three items, including **Run Oresme.** Selecting **Run Oresme** will produce a screen that looks similar to this:



You can select one or more concepts in the left window by clicking on them.   This effectively activates, or “turns on” the concepts. Clicking a second time turns them off.   Pressing the **Cycle Input** button shows what other concepts are activated by those concepts**.** The **Cycle Output** button does the same thing that the **Cycle Input** button does, except that it cycles the network output window back into itself. That is, instead of "thinking" about the concepts you originally gave it, it is thinking about the concepts generated by the concepts you originally gave it.

If there are too many concepts to fit in the dendogram, scroll bars are activated on the right. You can use the scroll bars to scroll through the rest of the concepts.



Network Input:
BEACHES
BIG
BROWN
BUFFALO
CITY
DETROIT
DIEGO
DROUGHT
EXCITING
FALLS
FRIENDLY
FUN
GOOD
LOTS
NEIGHBORS
NIAGARA
PALM
PARTY
SAN
SAND
SUN
SURF
THERE'S

Network Output:
BIG
BROWN
BUFFALO
CITY
DETROIT
DIEGO
DROUGHT
EXCITING
FALLS
FRIENDLY
FUN
GOOD
LOTS
NEIGHBORS
NIAGARA
PALM
PARTY
SAN
SAND
SUN
SURF
THERE'S
TREES

Here you'll notice that the concept "sand" has been highlighted in the input. When "cycle input" is selected, this causes Oresme to find the associations listed in the right or output window, *San Diego*, *Palm Trees*, *Sun*, *Sand* and *Surf*, which Oresme "knows" are associated because of the text it read.

If Oresme is giving you too many associations, you might try raising the threshold level (See Chip-head network options). Or, if you are getting no associations or too few, try lowering the threshold.

Oresme is a form of non-hierarchical cluster analysis. It is non-hierarchical because a) it doesn't distinguish between the name of a category and its members, but treats every concept alike, and b) it doesn't place each object into one and only one category, but can place any object into as many different categories as it belongs.

Oresme lets you enter the name of a category, and Oresme will then tell you what objects belong in that category. Or, you can name one or more items in a category, and Oresme will tell you the name of the category and the other objects in it. In general, Oresme answers this question: what other objects are in a category which includes x_i, x_2,...x_n, where x_1, x_2 and x_n are objects you selected.

Oresme is extremely useful for problems of this type: if a customer has bought items x, y and z, what other items might he or she buy? Or, (the same problem) if a customer who bought a new Ford Taurus was 46 years old, male, and had three children, what are the rest of his most likely demographic characteristics?

Chapter 4 Perceptual Maps

The dendograms are good for identifying clusters of concepts, but to get a better picture of the concepts you might like to look at conceptual maps. Figure 1 is an example of a perceptual map pasted from **ThoughtView**:

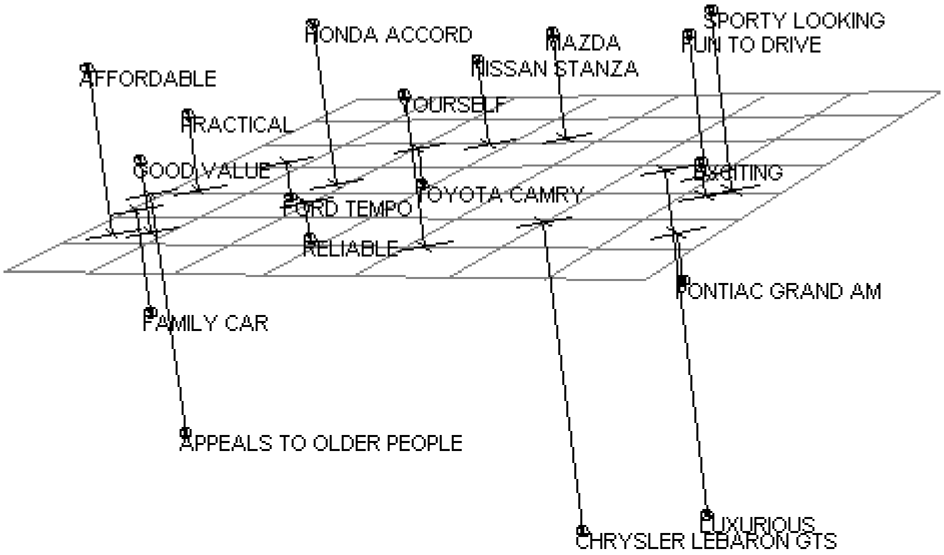
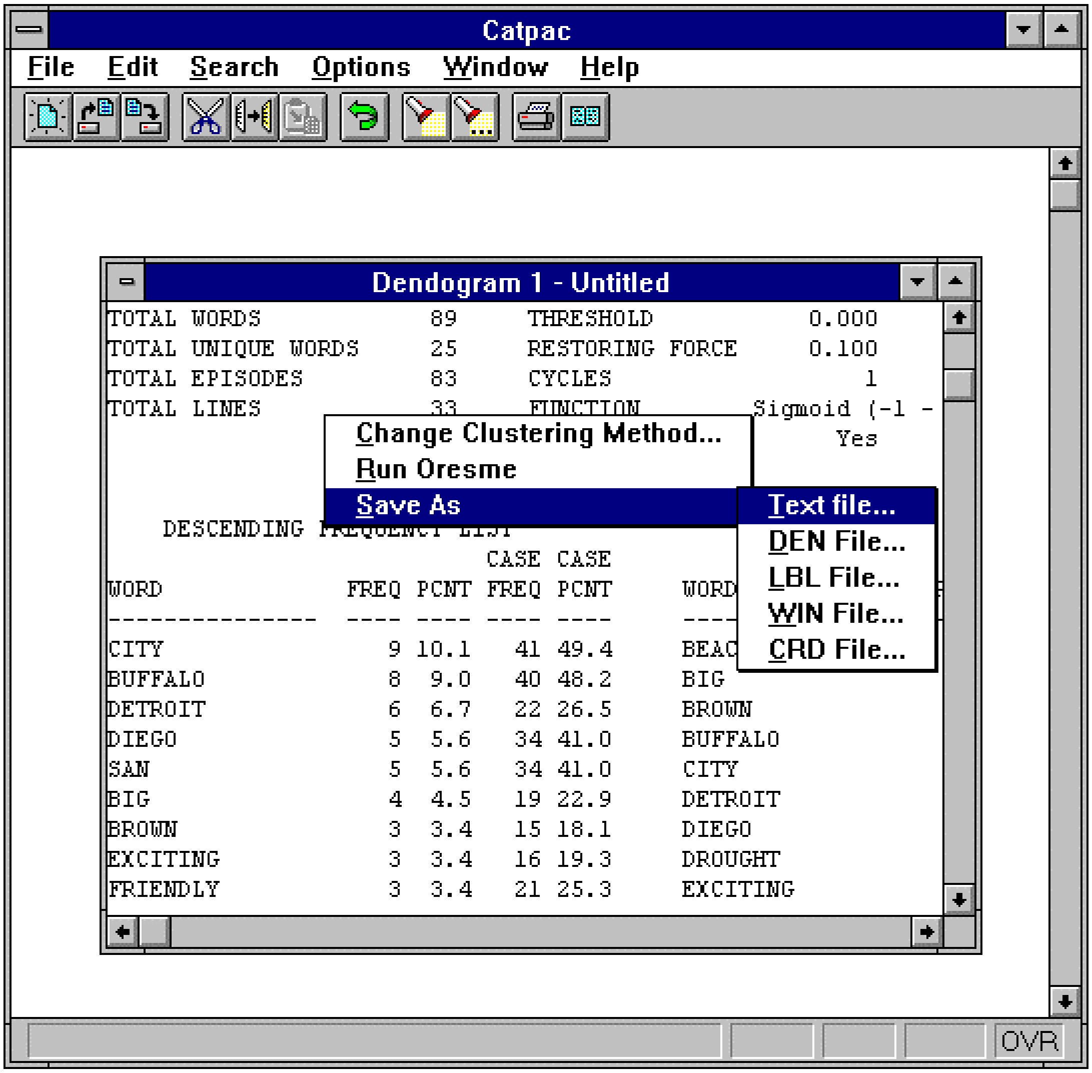


Figure 1 Perceptual Map

ThoughtView displays conceptual maps from conceptual map files, also known as a coordinate files, or ***.crd*** files -- pronounced "crud." **Catpac**'s Dendogram windows contain all the information necessary to produce a conceptual map file. The easiest way to produce a ***.crd*** file is to position the mouse on top of the window that contains the dendogram and click the right mouse button. This will produce a speed menu that contains three items, including **Save As.** Choosing **Save As** produces a second pop-up menu with several categories, including **CRD File** (See Figure 2). Select

CRD File to extract coordinate information from the dendogram window and place it in a ***.crd*** file. **ThoughtView** can read a ***.crd*** file and display it in either 2D, 3D, or stereo 3D.



```
TOTAL WORDS              89     THRESHOLD          0.000
TOTAL UNIQUE WORDS       25     RESTORING FORCE    0.100
TOTAL EPISODES           83     CYCLES                 1
TOTAL LINES              33     FUNCTION       Sigmoid (-1 -
                                                     Yes


   DESCENDING FREQUENCY LIST
                               CASE CASE
WORD               FREQ PCNT FREQ PCNT      WORD
----------------   ---- ---- ---- ----      ----
CITY                  9 10.1   41 49.4      BEAC
BUFFALO               8  9.0   40 48.2      BIG
DETROIT               6  6.7   22 26.5      BROWN
DIEGO                 5  5.6   34 41.0      BUFFALO
SAN                   5  5.6   34 41.0      CITY
BIG                   4  4.5   19 22.9      DETROIT
BROWN                 3  3.4   15 18.1      DIEGO
EXCITING              3  3.4   16 19.3      DROUGHT
FRIENDLY              3  3.4   21 25.3      EXCITING
```

Figure 2 Save As Speed Menu

Selecting Analysis Words

Catpac chooses words based on their frequency in the data file. The most frequent words are chosen first -- up to any limit you set in the **Analysis Words** section of the **Network Options** dialog box. You can use exclude files to force your analysis not to use certain words regardless of their frequency. You can also use include files to force words into the analysis regardless of their frequency.

Using an Exclude File

Many of the most frequent words are not content-bearing words at all. Such words include determiners, prepositions, etc. Catpac can read from an ASCII file that contains a list of these words, and exclude them from your analysis. By default Catpac reads a file called ***exclude.exc*** at startup. You can use a different exclude file by selecting **Open|Exclude File** from the **File** menu. By default Catpac only searches for files with the ***.exc*** extension. You can also create a new exclude file by selecting **New|Exclude File**. To edit the current exclude file double click on the icon.

Using an Include File

Sometimes some of the words you may be interested in don’t occur frequently enough to be included in your analysis. For example, let’s say you analyzed texts about pizza and limited the analysis to only ten words. You might find that the word pizza wasn’t one of the ten most frequent words and therefore wasn’t included in your analysis. However, if you use an include file containing the word *pizza*, it will be included in the analysis regardless of how frequently it was used. So in addition to the ten most frequent words, you will now have an eleventh -- pizza.

You can create a new include file by selecting **New|Include File** from Catpac’s File menu, or you can open an old include file by selecting **Open|Include File**. By default Catpac only searches for files with the ***.inc*** extension.

You can type as many include words as you like into the include file. Catpac will continue to use the include file until you close it. It will still use it if you minimize it. When minimized the include file looks like this: INC. Catpac does not use an include file by default, and only uses one include file at a time. Opening a new include file closes any current one that may be open.

Dendogram Windows

Dendogram windows, when minimized, look like this: . When you look at their contents, what you see is text in the form of frequency statistics, and a textual dendogram. But the window contains more complete information on your analysis. This additional information includes network parameters. So if you save a dendogram window you can re-read it with **Catpac** later, use **Oresme** with it, produce ***.crd*** files with it, etc. However, all this additional information makes no sense to a normal editor or word processor. If you want to save the text portion of the dendogram in a form that your word processor can read, you should save it as a text file. You can do this by choosing **Save As|Text File...** from the **File** menu. The extension ***.txt*** is added automatically to the file name you use.

Understanding Dendograms

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.

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 sub-cluster “you,” “want,” and “faster.” “Little” and “caesar” end up in a cluster that includes “little,” “caesar,” “inexpensive,” and “place,” with the sub-cluster “two” and “one” (Little Caesar’s offers two-for-one pizzas).

```
           CENTROID METHOD

G P D F D Y W F L P Q T O L C I P
O I O A E O A A I I U W N I A N L
O Z M S L U N S K Z A O E T E E A
D Z I T I . T T E Z L . . T S X C
. A N . V . . E . A I . . L A P E
. . O . E . . R . H T . . E R E .
. . . . R . . . . U Y . . . . N .
. . . . Y . . . . T . . . . . S .
. . . . . . . . . . . . . . . I .
. . . . . . . . . . . . . . . V .
. . . . . . . . . . . . . . . E .
. . . . . . . . . . . . . . . . .
. . . . . . . . . . . . . . . . .
. . . . . . . . . . . . . . . . .
. . . . . . . . . . . . . . . . .
. . . . . . . . . . . . . ^^^ . .
. . ^^^ . . . . . . . . . ^^^ . .
. . ^^^ . . . . . . . . . ^^^^^ .
. . ^^^ . . . . . . . . ^^^ ^^^^^ .
. . ^^^^^ . . . . . . ^^^ ^^^^^ .
. . ^^^^^ ^^^ . . . . ^^^ ^^^^^ .
. . ^^^^^ ^^^ . . ^^^ ^^^ ^^^^^ .
. . ^^^^^ ^^^^^ . ^^^ ^^^ ^^^^^ .
^^^ ^^^^^ ^^^^^ . ^^^ ^^^ ^^^^^ .
^^^ ^^^^^ ^^^^^ . ^^^ ^^^ ^^^^^^^
^^^ ^^^^^ ^^^^^ ^^^^^ ^^^ ^^^^^^^
^^^ ^^^^^ ^^^^^ ^^^^^ ^^^^^^^^^^^
^^^ ^^^^^^^^^^^ ^^^^^ ^^^^^^^^^^^
^^^ ^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^
^^^^^^^^^^^^^^^ ^^^^^^^^^^^^^^^^^
^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
```

Figure 3 Pizza Dendogram

Chapter 5 Understanding Frequency Statistics

Analysts often use the frequency information provided by **Catpac** to see which concepts occur most frequently in their data. They also use frequency information to help them see if their data needs any cleaning-up before making a second run. The frequency information can help you find typographical errors, synonyms, plurals, pro-nouns, and other such words that you may want to re-code. You can do this with **Catpac**'s **Replace** command found in the **Search** menu.

The following are the frequency statistics from a sample **Catpac** run. 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:

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

```
    DESCENDING FREQUENCY LIST                ALPHABETICALLY SORTED LIST

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

Figure 4 Frequency Statistics

The first group of statistics includes threshold, restoring force, cycles, function, and clamping. These are simply a record of the network parameters for this run. "TOTAL WORDS" is the total number of words in the text. "TOTAL UNIQUE WORDS" is the number of words used in the analysis. "TOTAL EPISODES" is the total number of windows used in the analysis. If you chose a case-delimited analysis, this is the number of cases. "TOTAL LINES" is the total number of lines in the text you analyzed.

The example in Figure 4 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 next group of statistics are for individual words, and they're sorted both alphabetically and by frequency. Under the column FREQ is the number of times that particular word occurred in the text. Under the word PCNT is the percentage of time that particular word was used in the text. For example, if a word has a PCNT of 5.0, and there are 200 total words in the text, than that particular word occurred 10 times.

The case frequency indicates the total number of windows in which a word was used. In case delimited data this is the number of cases that used that word. If you want to know how many people in your case delimited sample used the word "mileage" this is where to look. Case percentage is the percentage of windows that contain a particular word. In case delimited mode this is the percentage of cases that used a particular word.

In Figure 4, 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, and that it occurred 13 times, which was 11.3% of all occurrences. "Little" appeared in 58 or 42.0% of the scanned windows.

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.

Chapter 6 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-ended survey questions, focus group transcripts, newspaper and magazine articles downloaded from a data base, comments left on a customer telephone hotline, and restaurant/hotel/airline comment cards.

Figure 5 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 that make a difference, instead of attributes that all the products might share.

> *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 5 Pizza Interviews

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!

Weight Input Networks (.win file)

As we've said, Catpac works by finding the connections among a set of neurons that represent 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. In general, the ***.win*** matrix is a square proximity or similarities matrix whose rows and columns are the words identified by Catpac and whose entries represent the similarity or degree of association between pairs of words. It differs from a covariance matrix or correlation matrix mainly in its normalization. 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 analysis software packages.

Clustering Methods

Catpac includes the seven most commonly used clustering techniques for representing clusters of concepts as dendograms. There is no “perfect” clustering technique and each of these techniques has its advantages and disadvantages. In fact, there never will be a perfect clustering method, since the optimum clusters in any case depend at least in part of the needs of the analyst as well as the nature of the data.

All of these methods are of the agglomerative variety. They all share the same basic algorithm except for the distance metric used.

Dendograms impose a hierarchical structure that may not be appropriate for you particular data set, and for your research questions. You may prefer to use Oresme if you need non-hierarchical clustering.

Ward's Clustering Method

Like the centroid method, Ward's method represents clusters by their central point. Instead of using a Euclidean distance measure, Ward's methods measure the distance as the increase in the total sum of squares that would result from clustering together two objects.

Ward's method typically favors small spherically shaped clusters of roughly equally size. This may be good if smaller concept groupings of concepts are of interest, but it may make it harder to identify large groups of concepts that belong together.

Centroid Method

The centroid method represents objects by their centroids (the center of gravity of the points). The distance metric used is the Euclidean distance between the centers of gravity.

The centroid method is a non-monotonic, which can be interpreted to mean that the elements of two clusters may be more distant than the clusters themselves.

The centroid method is not sensitive to outliers. This means that if an outlier joins a cluster it will not influence the centroid of a cluster very much. Usually this is what is desired. This is a disadvantage if long chains of concepts are desired.

Median method

The median method is a variation on the centroid method, except that even weight is given to both clusters in determining the new centroid. This increases sensitivity to outliers and makes chaining more likely.

Like the centroid method, the median method is non-monotonic.

Single Linkage Method

In the single linkage method the distance between any two clusters is the distance between the two closest elements of the two clusters. This is also commonly referred to as the "nearest neighbor" technique

The single linkage method is known clustering in long chains, where the initial element of the cluster may be very different from the final element of the cluster. This method is ill-regarded because of this tendency.

The single linkage method is a very fast technique, and has been favored for clustering very large data sets because of this. It also does not place heavy demands on system memory. However, all the techniques used by Catpac use the most efficient algorithms, and time and memory should not pose a problem.

Complete Linkage Method

Also called the "farthest neighbor" technique, the distance between any two clusters is the distance between the two farthest elements of the two clusters.

Similar to the single linkage in efficiency, the complete linkage method does not suffer from chaining. It is actually biased against any sort of chaining. It will tend not to find clusters that contain two unrelated words that are both highly related to a third.

Group Average Method

The distance between two clusters is the average of the distances between every pair of objects in opposing clusters.

Weighted Group Average Method

The difference between the weighted group average, and the group average methods is fairly involved. But basically the difference is that the weighted group average gives equal weight to the smaller of two clusters upon their joining. Thus the smaller of the two clusters has equal influence on further clusterings. This may tend to produce longer chains of clusters.

Agglomerative Clustering Methods

As mentioned earlier, all the clustering techniques used in Catpac are agglomerative. Agglomerative clustering methods are methods that work by progressively grouping the most similar pair of objects until there is only one object left. They obey the following algorithm:

1. Calculate an N x N Distance matrix of the objects to be clustered.

2. Search for the smallest distance in a matrix.

3. Calculate a N-1 x N-1 using an appropriately chosen distance metric.

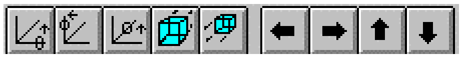
4. Repeat step 2 and 3 until only one object remains.

Chapter 7 Using ThoughtView

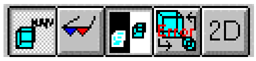
Chapter 8 Opening a coordinate (.crd) file

To open a coordinate file, first press the input file button, which looks like this: . Alternatively, you can select **Open** from the **ThoughtView File** menu. Either of these actions will bring up a standard dialog box, from which you can select any files that have the ***.crd*** extension.

Rotating, growing, etc.

You can rotate, shrink, grow, and otherwise move and manipulate coordinate spaces by pressing various buttons on the screen.   The nine movement buttons look like this: . The first three rotate around the x, y, and z axis respectively.   The fourth and fifth buttons move the space toward you and away from you respectively. The next four move the space to the left, right, up, and down.   To the right of these buttons are the shrink, grow, and center buttons:

.   The shrink button shrinks the space smaller. The grow button will make the space bigger. Notice that this is not exactly the same as what the forward and backward buttons do.   The center button will center the space in its window.

The next five buttons, 

, alter the display format of the space. The first button, which is pressed by default, toggles the names on and off.   The second changes between mono 3D and stereo 3D, which requires red and blue glasses. The third button flips between a black and a white background.

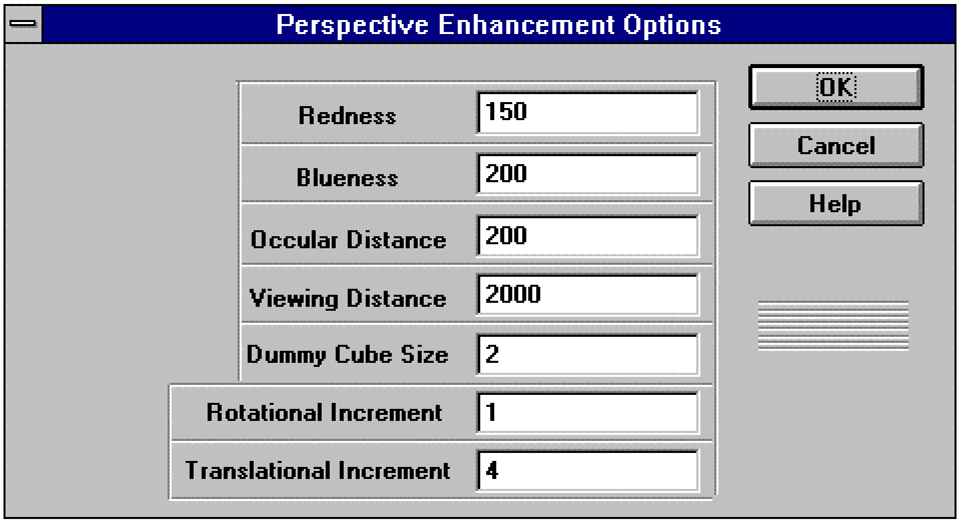
The fourth button is a little more difficult to explain. It toggles between the default cube size, which you can alter for best viewing, and a size that represents the standard error of the cube. For more information on this, see the section entitled “Default radii supplied”. The fifth button shows a more traditional two dimensional plot.

Perspective enhancement

Truly convincing 3D stereo perspective depends on many things. One of the most important is the relationship between the size and resolution of the viewing device, the viewer’s ocular distance (distance between the eyes), and the viewer’s viewing distance. Unfortunately, our software is not intelligent enough to figure out what size monitor you’re likely to have, nor how far you are sitting from it at any given moment.

Adjusting these is actually easier than you might think, and you can do it in a couple of minutes if you have a hand calculator handy[4]. If you select **Perspective Enhancement** from the **Options** menu you will get a dialog box that looks very similar to the following one:

[4] You do have one handy. It’s in your accessories group under the program manager.



Perspective Enhancement Options

Redness	150
Blueness	200
Occular Distance	200
Viewing Distance	2000
Dummy Cube Size	2
Rotational Increment	1
Translational Increment	4

If you truly desire ideal perspective, you should calculate how many pixels apart your eyes are, and how many pixels away from the screen you are. For example, if you're operating in standard VGA mode your resolution is 640 wide by 480 tall (the height isn't so important). If your monitor is 15 inches wide[5], then 640/15 equals 42.67 pixels per inch. From pupil center to pupil center, most people's eyes are about 2 1/2 inches apart. Multiplying 2 1/2 x 42.67 equals an ocular distance of 107 pixels. If you're sitting 2 1/2 screen widths away from the monitor, 2 1/2 x 640 equals a viewing distance of 1600 pixels.

Not all monitors display colors equally well, or at equal intensity. You should make sure that when you look through the blue lens of your glasses you can only see the blue lines when the 👓 button is pressed. Similarly, when you look through the red lens you should only see the red lines. If not, adjust the redness and blueness levels in the **Perspective Enhancement** dialog box until this is so. You can adjust the levels between 0 (darkest) and 255 (brightest).

[5] Note that when the salesperson says he/she is selling you a fifteen inch monitor, they are usually measuring the diagonal, not the width. You should measure the width with a ruler.

The defaults are set to work well with monitors that display colors accurately (I use a NEC), are about 10 1/2 inches wide, and are set to a resolution of 1027 width by 768 height. These are the defaults partly because this is a common setup, and partly because it's what I use most often.

Large Projection Screens

If you're in a hurry, try setting the ocular distance to 30. The rest of the default settings usually don't require much adjustment. You might also check the redness and blueness because many projection screens do a terrible job of reproducing color.

If you have a few minutes to spare, and you are a perfectionist, all the above formulae still apply.

Copying images

Full-motion 3D presentations are a lot of fun, but you still need to copy pictures into a report now and then. To copy the contents of the currently active window into your favorite word processor, press the copy button, which looks like this: . You can then switch to your word processor, paint program, or similar program, and paste the image as a bitmap.

The size of the image depends on the size of the window it was in when you pressed the copy button. If you want to change the size of the image, just change the size of the window.

The resolution of the bitmap depends on the resolution of your screen. If you find that the resolution of the bitmap is too grainy, you may want to switch to a higher resolution and re-copy the image.

Viewing multiple coordinate files

Like **Catpac**, **ThoughtView** uses a multiple document interface. As long as you have memory to spare you can look at as many ***.crd*** files as you like. This is especially useful if you’ve split your data into different demographic groups and have made ***.crd***s of each.

For example, suppose you want to compare males’ and females’ views on a particular topic[6]. Just open each of the ***.crd*** files like you normally would, and then select **Tile** from the **Window** menu. Figure 6 Comparison of Perceptual Maps is an example of males’ verses females’ views on the differences between men and women:

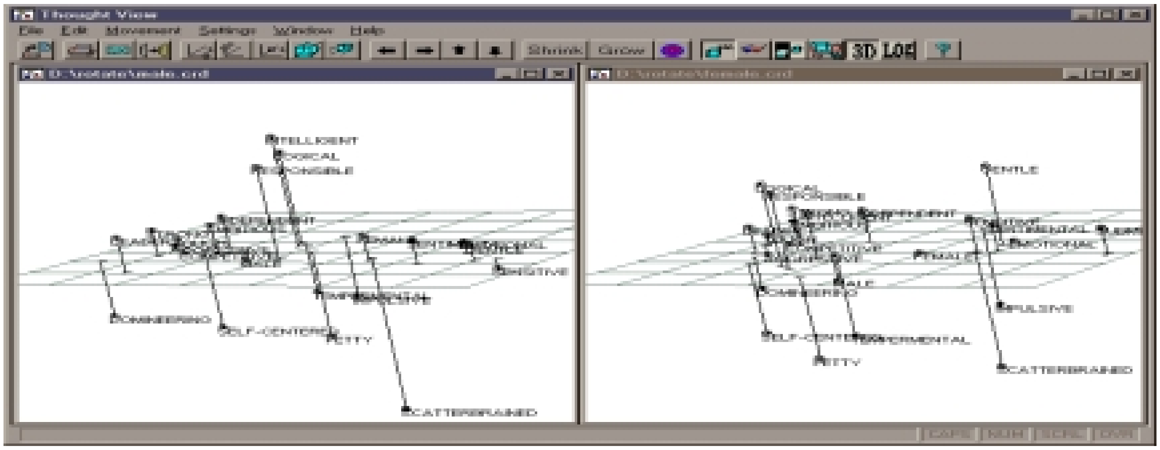


Figure 6 Comparison of Perceptual Maps

[6] When comparing multiple coordinate files, be careful that they share a common orientation; that is, they are all pointing in the same direction. Remember that the orientation of all perceptual maps is essentially arbitrary, and something needs to be done to assure they are the same from one map to another. The only way to assure this is through the use of a first-class “rotator” like **Galileo***ROTATE. ROTATE will take a series of sets of coordinates as input and rotate them until they share a common orientation, then write out the rotated coordinates for viewing with a viewer like **ThoughtView**.

The discerning viewer might notice that these two spaces don't look too similar. Are men and women really this different in their views? Actually, they're not, but they appear to be quite different, because the maps differ in their orientation. Perceptual mapping programs don't obey any conventions in deciding what part of the picture ought to be the top, bottom, left, right, or so. Before you compare two or more spaces, you should run your coordinates through a program which lines them up in a common orientation.

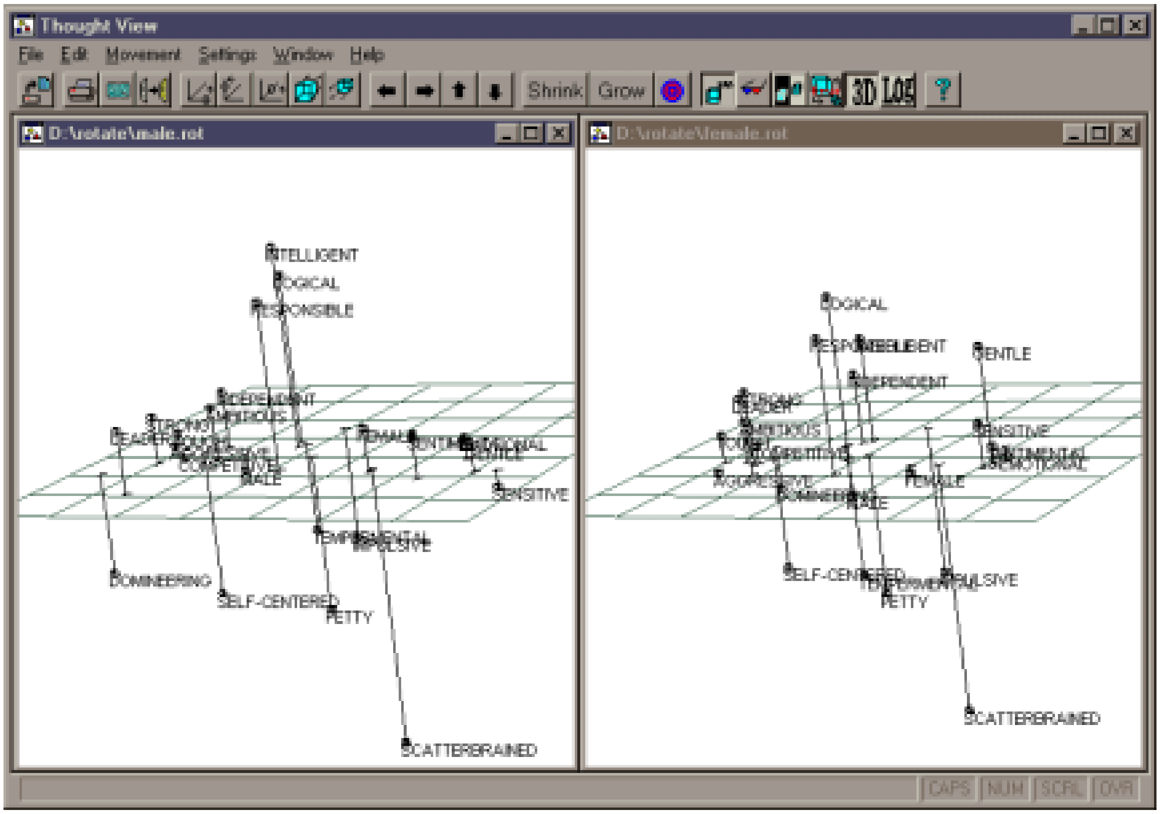


Figure 7 Gender spaces after rotation

Figure 7 compares the perceptual spaces of men and woman after the maps have been rotated to best fit on each other by Galileo*ROTATE^{tm}. Notice that the differences are much smaller than they appear in Figure 6. As a rule of thumb, spaces ought never be compared until after they have been reoriented by a rotation program like Galileo*ROTATE. (If you don't have such a program now, you can get Galileo*ROTATE from The Galileo Company.)

Default radii supplied?

You may have noticed that sometimes when you read a coordinate file, **ThoughtView** mentions that error information is not available, and that default radii are supplied. Well, no measuring instrument is perfectly accurate, and the people you're measuring don't always agree completely. Some measuring techniques can indicate the amount of error that surrounds the concepts in a space. **ThoughtView** visually represents these as either cubes or circles. If your coordinate file was produced by a technique capable of estimating these errors you can see them by pressing the "show 'true' cube size" button, which looks like this: Error. When you do this you can be reasonably sure that the concept actually lies somewhere within the cube. If the errors are missing from your coordinate file, **ThoughtView** will make arbitrary but attractive cubes and print the informative message and tell you that default radii have been supplied.

We do not currently have a way of estimating the error using **Catpac**. However, the **Galileo**™ software is capable of making such estimates.

Input to ThoughtView

Catpac can save your dendogram files as coordinate files (***.crd***) for subsequent reading by **ThoughtView** . If you would like to enable your own software to write coordinate files for **ThoughtView** we provide the following description of the ***.crd*** file format:

ThoughtView reads coordinate files in any standard ASCII format. The program expects to find four elements in the coordinate file: the *format header, the coordinates, the concept labels, and the standard errors.*

The format header

The format header is a standard format statement, followed by three 3-digit integers which give, respectively, the number of concepts, the number of real dimensions, and the total number of dimensions. Following these numbers a title up to 40 characters may follow:

(8f10.4) 018012018 Auto Test Data .

This header says that the coordinates will be found in (8f10.4) format, that there are 18 concepts, 12 real dimensions and 18 total dimensions in the space, and that the title of these coordinates is "Auto Test Data."

The Coordinates

Following the header come the actual coordinates, in the format described in the header.

The Concept Labels

Following the coordinates, **ThoughtView** expects to find the concept labels, one per line, in A40 format.

The Errors

Several **Galileo** programs are able to calculate confidence intervals around the location of points in the perceptual map. If these are available, the are arrayed following the concept labels in (11F7.3) format. **ThoughtView** uses these errors to determine the size of the circle used to represent the concept's position. If these errors are not present, **ThoughtView** provides default values.

A Sample Coordinate File

```
(6F12.4)   7  4  7DESERT PREFERENCES
     56.1659     -4.6299     -6.3528      3.1023       .9094       .7301
     20.6566
    -49.8766     -1.8973      3.0467      9.7793      -.8078     -8.6278
      4.2251
      1.2211    -18.0480      3.1706     -7.7617       .0200    -10.0956
     -5.5885
     28.5394      1.0056     17.5189      2.9665       .4623      9.2294
    -11.2909
    -40.7849      8.4968      3.6503     -7.3565      -.6609      8.0289
     15.8414
     20.6520     20.3833     -4.9479     -2.0051       .3342    -10.4801
     -9.6510
    -15.9168     -5.3104    -16.0858      1.2752      -.2576     11.2150
    -14.1927
HOT
COLD
SWEET
CHERRY PIE
ICE CREAM
TOM
BECKY
  4.537  4.360  3.889  3.005  1.827  2.593  3.359
```

Chapter 9 Technical Notes for Chip-Heads

RAM

Catpac and ThoughtView make very efficient use of RAM. They will actually share much of their executable code in the form of DLLs when run at the same time, so the amount of RAM used is less than the sum of the amount of RAM required for both programs. The amount of RAM required by Catpac is proportional to the square of the number of words in your analysis. The data file is also stored in main memory for faster access. Catpac will use virtual memory if necessary, but this can slow things down greatly.

Capacity & Speed

Catpac for Windows 3.1 can do analyses of up to about 160 words. Catpac for Windows NT has no limitation to the amount of words it can use in the analysis. The speed of analysis is nearly[7] proportional to the square of the words used in the analysis and linear to the size of your data file. Because it runs on a more efficient operating system, Catpac for Windows NT is twice as fast as Catpac for Windows 3.1

Due to the nature of the full motion graphics it uses, ThoughtView requires a lot of processing power even when idle. So for optimal speed you may be better off running Catpac (or other processor

[7] The matrix routines in Catpac are designed such that they are more efficient as they grow larger. Otherwise it would be exactly proportional to the square of the words.

intensive applications) without running ThoughtView in the background.

Some Final Remarks

Catpac™ represents a new generation of artificial neural software that can do things older 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 Galileo representative.