INFORMATION PROCESSING 74 – NORTH-HOLLAND PUBLISHING COMPANY (1974)

COMPUTER GENERATED ANALOGUES OF MENTAL STRUCTURES FROM LANGUAGE DATA*

John B. SMITH

The Computation Center and The English Department
The Pennsylvania State University
University Park, Pennsylvania, USA

CGAMS produces a comprehensive image of the associative relations among thematic groups of natural language words. It displays on a t.v.-like graphics system an image resembling an aerial photo of a mountain range that reflects relative prominence of themes, their relative closeness to one another, and an indication of whether this relation is approximately stable or whether it varies in the text. The system has a variety of options enabling the user to explore his data in real time in close detail while retaining a sense of total pattern. Because the system assumes only a lattice model of data, it has potential applications in a variety of non-language areas.

1. INTRODUCTION

The project described below grew out of a continuing study of James Joyce's A Portrait of the Artist as a Young Man that was described at IFIP '71 in Ljubljana. The earlier study attempted to show that 1) major moments (epiphanies, as Joyce calls them) in the development of Stephen Dedalus' personality are accompanied by large concentrations of important images and 2) since Stephen's personality tends to be stable between these points, the changing structure of his mind can be traced by first describing the patterns of associations among images and then noting the changes in these patterns that take place at epiphanal moments.

In that study (see "A Computational Analysis of Imagery in James Joyce's A Portrait of the Artist as a Young Man," Proceedings of the IFIP '71 Conference, The Hague: North Holland Publishing Co., 1972, pp. 1443-47) an exact model that produced a numerical value reflecting richness of imagery as a function of number and importance of the images present in a section of text was proposed, and it demonstrated the validity of the first hypothesis--that moments of major change are accompanied by large concentrations of important images. To trace the changing patterns of associations among images--associativity between two images was taken to mean that two images are "close" to one another in the text viewed as a linear sequence of words--was more difficult. The entire sequence of associative relations (some thirteen thousand relations for Portrait) can be viewed as a lattice, or more properly, a ring structure. Although the computer can store and manipulate such structures easily, it is virtually impossible for the researcher--literary critic in this instance--to grasp this entire complex comprehensively. To aid in this task a number of computer aids were produced including an image concordance, distributions of thematic groups of images, frequency lists by chapter, and statistically defined clusters of images. With these aids, I was able to demonstrate the basic validity of the second hypothesis but not with the same exactness that the well-defined model afforded with respect to the first.

Since that time I have sought a model for associative structure that is exact and that allows a comprehensive view of the entire structure while containing detailed information for close, extended scrutiny of specific relations. The system that will be described, in my opinion, realizes these goals. Since it is quite general--it can be used with any language sample or, for that matter, with any data that can be viewed as a linear sequence of discernable entities or states--the system has potential application in a number of fields other than linguistics or literary studies. Below, I shall discuss the functions and use of the system, the computational procedures involved, and finally some areas where it might be helpful that may not be immediately obvious. For convenience, the system will be described as it is used for a thematic study of Joyce's Portrait.

2. CGAMS IMAGE

Since the model reflects basic patterns of associations among thematic groups of words, it may be regarded as a model of mind to the extent that the designated themes represent areas of major concern for a character or individual; hence, the system has been designated Computer Generated Analogues of Mental Structures and will be referred to by the acronym, CGAMS. The CGAMS image can be generated from any language text or designated subsection and is displayed on an ADAGE graphics terminal--a t.v.-like device with considerable computational support. The basic picture (see figure 1) contains a peak for each thematic group of words or images. The height of a peak reflects the prevalence or frequency of the thematic group for the section of text under consideration. The distance between two peaks reflects the tendency of the two themes to be close to one another or far apart, relative to their relations with the other themes. The facets of the peaks on a line between them indicate whether the two themes tend to be a stable distance apart (resulting in a sharp facet) or whether they are sometimes close and sometimes far apart (resulting in a sloping facet). In figure 1, themes designated A, B, C, and F form a central, interactive group; themes D and E are relatively uninvolved. Theme F is the most prevalent or frequent of the six, followed closely by B and C;

*This project was made possible by a grant for computer studies in the humanities from the American Council of Learned Societies. I am also grateful for support in the form of computer time from the Pennsylvania State University and for assistance in coding the computer procedures from Jay Gibbons and Paul Schuepp.

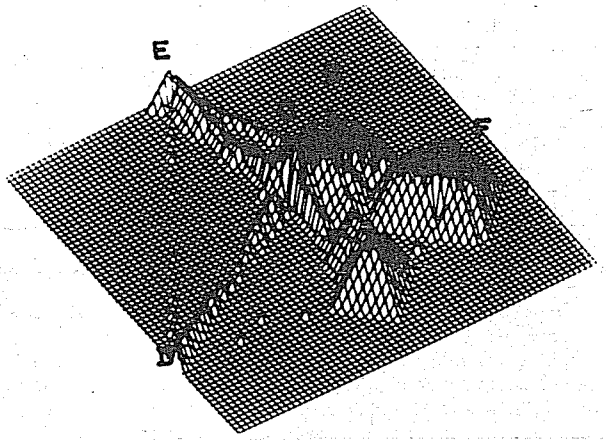

Figure 1
Basic CGAMS Image

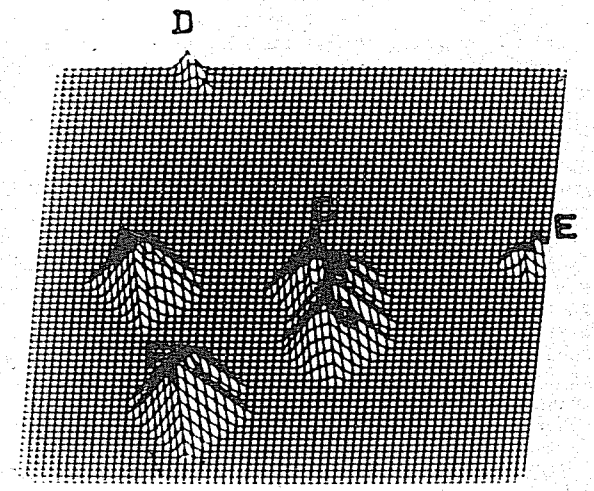

Figure 2
CGAMS Image, Horizontal Scale Reduced

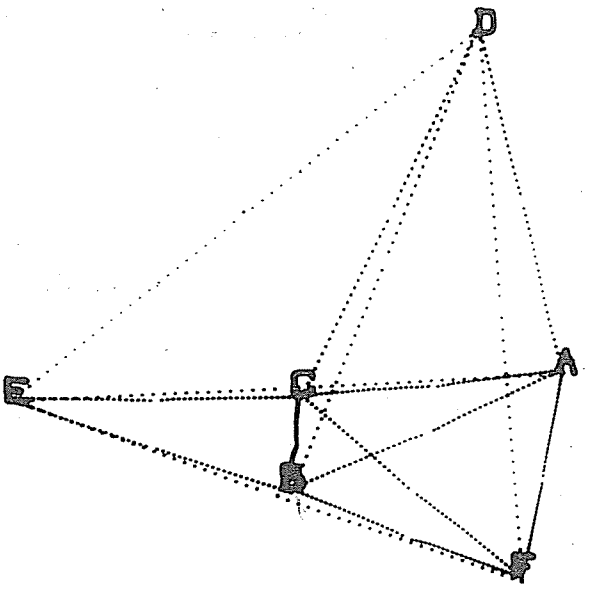

Figure 3
CGAMS Image of Raw Data Points, Top View

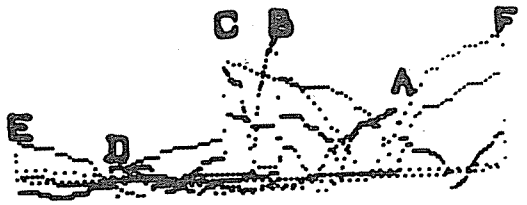

Figure 4
CGAMS Image of Raw Data Points, Side View

D and E are the lowest in frequency and, hence, the smallest in size.

In addition to generating the basic picture, CGAMS offers the researcher a number of options that alter the image for different effects. During the image generation stage, the researcher may request that the image be smoothed. This feature broadens the base of high peaks emphasizing relative height from virtually all viewing perspectives. Another option that may also be exercised at this stage allows for scaling in the horizontal plane. The default condition results in contours that meet at the midpoints between peaks. At the user's option, these contours may occupy any fraction of the distance between peaks; for example, a scale factor of four produces contours that cover one-fourth of the distance making the peaks distinctly separate and emphasizing height and relative horizontal distance but reducing definition of the various facets. Alternatively, all contours can be scaled according to some fraction of the minimum distance separating any two peaks. This option results in contours that descend over a uniform absolute distance producing an image in which peaks tend to be distinctly separate (see figure 2).

Once the image is projected by the ADAGE graphics system, a number of additional options are available for interactive, real time analysis. By turning one of three control knobs, the user may rotate the image around its x, y, or z axis. Similarly a Joy stick, a vertical lever like the control stick of an airplane, may be used to rotate the image obliquely in more than one dimension. By pressing a button, the researcher may add labels directly above the peaks; a second press of the button removes them. Another function displays a legend in which the symbols are interpreted. A third slowly rotates the image around the y axis. A fourth removes the grid lines and exposes the raw data points of the contours themselves and, thus, facilitates close, extended scrutiny of the relations represented (see figures 3 and 4). As with the basic image, labels and legend may be applied to this raw data image. A final function button produces a hard copy on the ADAGE printer-plotter of whatever image is currently displayed on the screen. Using these options, the researcher may sit at the ADAGE console and explore his data from a variety of viewing perspectives, refer back to the raw data from which the image is constructed, obtain a copy for later reference, call up similar images for other thematic groups or sections of text; in short, he has the resources to explore the text, creatively, comprehensively, yet also in detail.

3. CGAMS SYSTEM

The system was designed to accept natural language data in a format nearly identical to normal type

setting conventions; however, CGAMS will accept any data that can be represented as a sequence of discrete symbols or combinations of symbols. Once the data is prepared in a machine readable form, there are three major stages of processing necessary to produce the CGAMS image described above.

The first stage accepts the raw data and produces a highly flexible, list-structured representation of it. For natural language texts, data is entered virtually as it would appear on a printed page except for several minor conventions. It is then scanned, each word extracted and numbered in several ways, and written onto magnetic tape. For data not conforming to CGAMS input specifications, the user may write or have written a simple scan program to create a data set conforming to the scanned output format and continue processing from there. Output records, one for each word in the text, are sorted and again written onto tape. Finally they are processed by the RATS system (see John B. Smith, "RATS: A Middle Level Language Utility," CHum, 6, 5, (May, 1972), 277-83) where the data is divided into a dictionary of word types and two files of index information, one ordered sequentially, the other alphabetically; all three files are linked into a ring structure through pointers. In this format, the text is readily available for a wide variety of analyses, including CGAMS.

The second processing stage computes, using a RATS representation of the text, the data for the CGAMS image. The user first constructs a system of categories or themes he wishes to display be selecting from the RATS dictionary the items that are to be categorized. This information is represented by a category or theme title followed by the numerical positions in the RATS dictionary of the words or symbols that constitute that category.

Using the RATS text and the list of designated themes, CGAMS then computes distributions of the distance relations among these themes. That is, it computes the number of times any word or symbol in category one is within one word or symbol of any member of category two, within two words, within three words, etc.; the process is repeated for all combinations of themes or categories. The user, at this stage, may specify the number of cells for each distribution and he may also specify a scale factor for each cell: for example, the first cell might represent distance relations of 1 to 10; the second, 11 to 20, etc. These distributions are stored on tape or disk for later processing.

From the distributions, a matrix of characteristic distances between each pair of themes is computed. Currently the user may select either the mean or the median. On the basis of these distance relations, CGAMS determines the position in the horizontal, xy plane where the peak for each theme will be located using the Kruskal-Shepherd nonmetric multidimensional scaling algorithm (see J. B. Kruskal, "Nonmetric Multidimensional Scaling: A Numerical Method," Psychometrika, 1964, 29, 115-29). This program locates two-dimensional coordinates for points, corresponding to the themes or categories, such that the relative distances among them reflect the relative characteristic distances produced in the preceding step. These coordinates are passed on to the third CGAMS program that adds the z, frequency dimension for the image to be displayed.

Using the coordinates computed by the scaling program for the center of each peak and the distributions of distance relations, CGAMS constructs two representations of the data. For each pair of themes, CGAMS computes the equation for the straight line between them. Starting at some point between them, it accumulates values in the distributions (the z coordinate values) and spaces these values uniformly along that line (the xy coordinate values) until the distribution peaks at the location determined by the scaling program--the center of the peak being constructed. The point where the distribution begins may be the center of the line between the two peaks, any fraction of that distance, or some absolute distance such as half the minimum distance between peaks. The data computed in this manner constitutes the raw data depicted in figures 3 and 4.

The mountain range effect is produced from this raw form of the data. A uniform grid is placed over the horizontal, xy plane; the user may specify the number of lines, but 70 is a practical limit on our ADAGE system. Each point in the raw data is interpolated to the nearest grid intersection and the maximum z value of all such interpolated values for a grid intersection retained. To enhance the vertical height of the resulting peaks, the user may specify a maximum z displacement; CGAMS will then make several passes through the grid data, effectively raising the floor a little each time to meet the actual data values minus the displacement. The base of each peak will thus be broadened one square for each pass through the data until no two adjacent grid points are more than the maximum displacement value from one another in height. Figure 1 illustrates the effect of this smoothing option.

The data computed in this step is then passed to the third stage, the actual display program. Note that all processing until this point is done on the main computer, independent of any specific graphics system. The user without an ADAGE display unit could use all of the preceding programs and then write his own display program for whatever graphics device he has available.

The third group of programs runs on the ADAGE computer and creates the visual image from the three-dimensional coordinates of the raw data and the grid data. It displays the points or constructs lines between them and by sampling data from the various analogue input devices provided by the ADAGE system--function switches, joy stick, buttons, etc.--moves the image in three space, produces a hard copy, adds labels, etc. With this display program running, the user may sit at the console and examine his data in real time.

4. APPLICATIONS

CGAMS has a wide range of potential literary and linguistic applications. So far, I have described the system primarily in the context of thematic analysis; this use can be extended by having CGAMS compute multiple images of a text. Separate images may be computed for, say, the first 1,000 words of a text, the first 2,000, the first 3,000, etc. cumulatively through the entire length. By displaying these various images, the user may use CGAMS to get a sense of the development of associative structure over a text. If a video tape or motion picture were produced from these cumulative images, one could see themes or categories grow and shift in their relations among one another over a novel or some other text. Perhaps soon such media products will become legitimate modes of literary criticism and take a position beside the monograph and journal article.

A number of other linguistic applications are possible. Other semantic groupings might be introduced and CGAMS used to monitor their relations. For example, there has long been interest in developing descriptive tools for characterizing thesaural structures. Sally Sedelow has developed

some useful measures in her ring structure VIA in the notion of distance and connectivity among thesaural categories (see Sally Y. Sedelow, et. al., Automated Analysis of Language Style and Structure in Technical and Other Documents, Contract #N 00014-70-A-0357-001, Office of Naval Research, 1971, Lawrence: The University of Kansas). CGAMS might be useful in displaying this information in a form that can be easily grasped intellectually.

Looking at other language features, CGAMS could be used to study relations among sounds or phonemes if a phonemic transcription were encoded instead of a graphemic representation. Similarly, syntactic categories might be encoded or derived later off the RATS dictionary and CGAMS used to characterize associative tendencies among these categories.

A number of other applications for literary and linguistics analyses could be listed; however, it may be useful to speculate concerning other language applications outside a literary or linguistic context. For example, it is widely recognized that psychological patients often reveal sources of deep-seated anxiety indirectly in their conversations. Relations among troubling experiences or themes may be reflected in patterns of continual proximity or association that are unrelated syntactically and which may be vehemently denied by the patient. The studies by Howard Iker and Norman Harway (see Howard P. Iker and Norman I. Harway, "A Computer System Approach towards the Recognition and Analysis of Content," Computer Studies in the Humanities, vol. 1, no. 3 (1968), 134-54) as well as others indicate that the computer can help in defining basic themes from transcripts of psychotherapy sessions. CGAMS could be used to extend this work by also revealing the fundamental patterns of relations among these themes. Hard copies of data from individual therapy sessions would make it possible for a doctor to monitor over a period of time changes too subtle to notice otherwise. When that day finally arrives when we have a speech recognition device, the system could be run in real time enabling continuous monitoring of a session or conversation. Such uses are promising for a number of fields other than psychiatry and linguistics.

Since CGAMS is completely general beyond the initial encoding stage, it may find applications in fields not directly related to linguistics or language study. The system assumes only that the data is a lattice involving nodes or symbols and paths or relations among them; consequently anything that may be considered a lattice could employ CGAMS to characterize its structure of relations. A major problem with large virtual computing systems is the degree of scattering of memory fetches and branching instructions. Programs that proceed sequentially through their code without branching out of the immediate program context and which, similarly, access data sequentially or from a restricted environment are best suited for such systems. With virtual systems, using paging, a large number of programs that do not exhibit local concentration of activity can bring the system almost to a standstill just swapping pages in and out of main storage. Since main storage, actual or virtual, can be considered a linear string or vector of page units, CGAMS could display the degree of scattering of branches and fetches for the system's recent past. Where allocation of resources such as partition and page size are under operator control, CGAMS could facilitate making these decisions as well as providing materials for demonstration purposes.

Although we don't usually think of the computer itself as an information retrieval system, many of the problems faced by the operating system are micro versions of the problems faced by any sophisticated conventional information retrieval system. Again the problem of dispersness is a major factor affecting performance and must be monitored. Certainly CGAMS would not replace precise statistical measures of operation; but, again, it could be used to present a conceptual indication of a system's overall organization and performance.

The list of applications could be extended almost indefinitely. Other obvious areas concern the design of microcircuits and monitoring of macrocircuits or networks and traffic flow; however, from these examples the reader can get some idea of the system's scope and generality. It is particularly exciting to me that problems growing out of literary analysis and linguistics when reduced to their bare structural form are the very same problems faced in a number of other disciplines. If humanists working on the problems that interest them can benefit as well as learn from those working in fields more obviously computational, we may look forward to computer applications that are humanistic in the true and full sense of that word.