Preprint of English version published in French as Rockwell, Geoffrey and John Bradley, "Empreintes dans le sable: Visualisation scientifique et analyse de texte", in *Litterature, informatique, lecture* edited by Vuillemin and LeNoble, Paris: Pulim, p. 130-160, 1999. French reprinted online at *L'Astrolabe* at http://www.uottawa.ca/academic/arts/astrolabe/.

Printing in Sand; Scientific Visualization and the Analysis of Texts[1]

"seeing is forgetting the name of the thing one sees"

Paul Valéry (1871-1945)

Geoffrey Rockwell
Assistant Professor of Humanities Computing and
Director, Humanities Computing Centre
McMaster University

John Bradley
Director, Media Services
University of Toronto

Introduction

The computer generated image of air flow over the wing of a jet plane in a wind tunnel and the rendering of the movement of subatomic particles during a chemical reaction are both examples of Scientific Visualization, a technique that has generated significant interest in scientific and engineering fields. Scientific visualizations usually represent vast amounts of quantitative data that would be too hard to read in an interactive and rhetorically effective form. But could these techniques be applied to the Humanities? More concretely, what does Scientific Visualization teach us about the graphical understanding of information in general, and the analysis of texts in particular? This chapter will look at the lessons to be learned from Scientific Visualization about the possibilities for the visualization of texts. It is our contention that:

- there is today in computer aided textual studies a body of standard techniques and data structures that could form the basis for a textual visualization package, and
- that this development could make such tools more widely available to those in the community who do not have the resources needed to create such tools themselves.

Furthermore, we believe that more work in this vein will allow computing tools to be developed that could be usefully applied to many research issues in the humanities, and might be used by a larger community in it. To do this we will:

1. first discuss Scientific Visualization,
2. then discuss the application of visualization techniques to the analysis of texts,

---

[1] This paper was based on a presentation given at the ACH-ALLC in Paris, 1994; Bradley, J. & Rockwell, G., "What Scientific Visualization Teaches Us about Text Analysis".

Preprint of English version published in French as Rockwell, Geoffrey and John Bradley, "Empreintes dans le sable: Visualisation scientifique et analyse de texte", in *Litterature, informatique, lecture* edited by Vuillemin and LeNoble, Paris: Pulim, p. 130-160, 1999. French reprinted online at *L'Astrolabe* at http://www.uottawa.ca/academic/arts/astrolabe/.

3. and finally conclude with some observations about visualization and reading.

Scientific Visualization

Our ability to gather and to generate data with computers is overwhelming our ability to interpret such data. It is common in the sciences to gather so much quantitative data that manual interpretation of results is almost impossible. Reading tables of numbers to find anomalies or patterns ceases to be useful after a certain point. This is one reason why, since the 18th century, we have used graphics to present quantitative information. William Playfair, one of the inventors of statistical graphics, pointed out over 200 years ago that:

> Information, that is imperfectly acquired, is generally as imperfectly retained; and a man who has carefully investigated a printed table, finds, when done, that he has only a very faint and partial idea of what he has read; and that like a figure imprinted on sand, is soon totally erased and defaced. ... On inspecting any one of these Charts attentively, a sufficiently distinct impression will be made, to remain unimpaired for a considerable time, and the idea which does remain will be simple and complete, at once including the duration and the amount.[2]

Well designed graphical presentations of data not only help the viewer remember the information, they also help the viewer quickly find patterns in the data and anomalies. Or, as Tufte puts it prescriptively in *The Visual Display of Quantitative Information*, "Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space."[3]

<Figure 1: Combination of Table on one side and graph on the other>

Visualization builds on the rhetorical effectiveness of graphical representations by turning images into exploratory tools. Graphs have traditionally been used to present information understood by a researcher to an audience of peers or the general public. The graphical representation of the data follows the understanding. Visualization, by contrast, uses the graphical representation to assist the understanding of data. Visualizations are used by the researcher herself to explore the data for patterns or analyze it for details. For this reason visualizations are not only used earlier in the research process, but they are also typically interactive. Visualizations, when they are generated by a computer, need not be static representations, they can be dynamic ones that the researcher can play with in order to explore the data. Visualizations on computers are tools for exploration as much as representations.

Although visualization models and methods are very general and applicable to many different scientific problems, part of their usefulness is that they can be applied to a *specific* problem that an individual engineer or scientist is facing. As a result of the packaging of these tools, it is now possible to use them with a minimum of programming, or extensive mathematical analysis. With minimal distraction, the researcher is able to see visual images that represent his data on the screen – to actually "visualize" his data. By seeing the data in this way the researcher is able to gain better insight into what structures and forces may lay

---

[2] Playfair, *The Commercial and Political Atlas*, p. 3-4, from Tufte, *The Visual Display of Quantitative Information*, p. 32.

[3] Tufte, *The Visual Display of Quantitative Information*, p. 51.

behind it. In some cases, patterns and abnormalities might be strikingly visible that were, even when the raw data was worked over with statistical methods, only partly perceived, or missed altogether.

What is the source of the rhetorical effectiveness of visualizations? One obvious cause for the striking impact of visualizations is the sense of the reality of the object. Obviously, the high-definition 3-D rendering of the image (on a 2-D computer screen) engages the eye, but there is more to it than that. The 3-D visual image is generated using computing technology that allows for its rapid rotation and expansion. The ability to rapidly rotate the object and to look at it from various directions makes the simulated object become a "real thing" that can be manipulated. The presentation of live direct-manipulation graphics liberates the natural intuition that we all have that helps us interact effectively with things that we see all the time in the natural world. It becomes engaged in the process of understanding. The graphical image, instead of being simply a presentation tool, becomes an important research instrument, most effective at the time that the researcher is still grappling with the basic structures of the data s/he has presented to the machine.

Underlying the acceptance of Scientific Visualization is a culture of interpretation. At least a part of the reason that Scientific Visualization has become so successful is that the data structures and methods that make it possible are widely accepted in scientific circles. In the sciences there is a tradition of using and teaching these methods so that visualizations that depend on complex methods are easily grasped. To be specific, the acceptance of Scientific Visualization tools in the sciences is the result of two forces:

1. The standardization and formalization of data structures and appropriate processing (in the case of Scientific Visualization, principally a certain branch of mathematics), and
2. A wide acceptance and understanding of the mathematical methods so that non-mathematicians, and non-programmers can usefully make use of these tools with much reduced time and effort.

It is due to this standardization and acceptance within the discipline that visualizations become meaningful and tools can be built with a broad application. A visualization can mean a thousand words, but it doesn't, without standards, necessarily mean the same thousand words for each viewer. With agreement regarding the structure of data and methods to be used, visualizations can be reliably interpreted and tools can be built for general use.

Visualization Tools

<Figure 2: Explorer Screen>

Part of the success of Scientific Visualization in the sciences is also due to the quality of the tools that are available. In the Scientific Visualization world, the packaging of standard tools has progressed in software packages such as AVL, Explorer, or Khorus to the point that they all resemble each other and all three can be productively used without needing to do any programming in the traditional sense. A genre of visualization tool has emerged that is easy to use and has certain common characteristics. Figure 2 shows a typical screen created by Explorer. The software here is showing a projection of some geographic data onto a map of Korea. R.A. Earnshaw and N. Wiseman's book, *An Introductory Guide to Scientific Visualization,* surveys the state of Scientific Visualization as it existed in 1992. In their discussion of Khorus, they identify 5 major components in this type of Scientific Visualization software, but all these components are also present in AVL and Explorer:

- **Interoperable Data Exchange**: A collection of data structures that represent the basic items of discourse within the system.

- **Data Processing Libraries**: collection of standard operators that work over the range of supported data types. Explorer gives the user a palette displaying the names of the various operators – shown in the window on the top left in Figure 2.
- **Renderers**: Software that takes resultant data and displays it on the screen. An Explorer renderer is shown in the small window near the top right in Figure 2.
- **A Visual Programming Environment**: This is a two dimensional space, rendered on the computer screen, where processing nodes are connected together to provide a flow of data. The Nodes are created from a library of operators.
- **User Interface Development System**: A programmer's assistant – a collection of routines that can be called by any programmer who is creating his/her own operators or graphic displays.

The first three of these components (we list them in a different order than the order used by Earnshaw and Wiseman in their book) show direct connections with ideas first expressed in the UNIX Operating System – including the provision of standard data types, a set of basic operators that can be applied to them, and rudimentary tools to connect them together. The development of Graphical User Interfaces provided a context for the introduction of a more sophisticated operating framework: the idea of a "workspace" – a two dimensional space on which the user assembled the available operators he or she wished to use and indicated how the data flowed between them. Finally, developments in programming paradigms – principally Object Oriented Programming – have made it possible to design pieces of software in such a way that data transformation components can be developed independently of the coding for the workspace itself – thus allowing others to augment the system, even after the workspace itself has been completed.

One potentially confusing aspect of this genre of visualization tools is the use of visual programming environments. Typical of these visualization tools is that they have an environment where the researcher can create the programs that process the data that create the interactive visualizations. These programs are not created by writing code, but by dragging components (that process the data) out onto a workspace and connecting them with pipes for the data to flow through until the data can be graphed in an appropriate way. The result is that the workplace is a visualization of another sort. It is a data-flow visualization that shows the logic of the process whereby the data is prepared for graphing, not a visualization of the data itself. It is a visualization of the process not the result, though it is not surprising that in a community interested in visualization there would emerge tools also for the visualization of process.

To summarize, Scientific Visualization has the following aspects:

- Visualization as a technique is designed to deal with large quantities of quantitative data, usually multivariate data, and often data that has some spatial or temporal dimensions.
- Visualizations are typically interactive computer-based graphical representations designed for exploring and understanding data, not just presenting it to others.
- Visualizations depend on a data standards and wide acceptance of methods in the community that uses them.
- Visualization tools often use graphical programming environments.

Applications of Scientific Visualization to Text Analysis

No one disagrees with the assertion that for scientific/engineering data which is mainly spatial, a suitable graphical representation is a natural and comfortable – even powerful –

vehicle. With a text, however, it is only rarely that we have any obvious, uncontroversial, reference to any three dimensional metric space. Does any part of the Scientific Visualization model have any applicability to textual studies? In this second section of this chapter, we will present a few issues that might help us to see what might be also appropriate for humanities research – to examine, in other words, what could be a basis upon which a set of "textual visualization" tools could be created.

Graphical Interpretation

Visualization techniques have been applied outside of the physical sciences. One place to start a examination of the uses of visualization in the study of texts is to look at related types of information visualization. One such area is the study of visual programming techniques such as those mentioned above for the visualization of the data-flow.[4] Visual programming environments allow the user to build a program by creating a "flow-chart" of the intended program instead of typing code. The visualization of the program replaces the traditional lines and lines of code.

<Figure 3: Prograph Screen>

One of the interesting problems that arises with visual programming environments is that "the attractions of graphical representations are not matched by performance".[5] Research has shown that graphical representations, while attractive, can mean very different things to different programmers. Much of the information conveyed in a visualization comes from what Marian Petre calls "secondary notation" – the layout of the objects in the chart, how close objects are, colour, typographic clues, and other enhancements. Such features do not have unambiguous interpretations; what, for example, does an icon being above another mean? Also, many of these graphical features are not precise; it is hard to say, at a glance, exactly which icon is the closest to another. Thus, where there is no tradition of interpretation, or a community that is educated to read such visualizations, graphical representations can lead to confusion. For this reason researchers have found that visual programming tends to work best in specific domains, like Scientific Visualization, where agreement can develop about the objects and methods.[6] In a specific domain the secondary notation can be formalized so that there are standards as to what these features mean. This is another way of saying that it is an issue of the standardization of data structures and methods. For there to be a tradition of interpretation there has to be agreement as to what is being interpreted and how it is being interpreted.

But is there such a community of interpretation in the humanities or among those that closely study texts? In Rosanne Potter's survey of the application of statistical methods to the study of texts in the 25th anniversary edition of *Computers in the Humanities* she noted that work was needed to formalize a number of suitable methods that could be made available to scholars without requiring them to become statistics experts. It is our contention that there is now a body of suitable data structures and techniques that could form the interpretative ground for the broader use of visualization in textual studies, but what evidence is there of this body?

Topology of Graphical Representations of Texts

---

[4] For an introduction to visual programming see the March 1995 issue of *IEEE Computer*.

[5] Petre, "Why Looking Isn't Always Seeing: Readership Skills and Graphical Programming", p. 35.

[6] See Repenning, "Agentsheets: A Medium for Creating Domain-Oriented Visual Languages".

Preprint of English version published in French as Rockwell, Geoffrey and John Bradley, "Empreintes dans le sable: Visualisation scientifique et analyse de texte", in *Litterature, informatique, lecture* edited by Vuillemin and LeNoble, Paris: Pulim, p. 130-160, 1999. French reprinted online at *L'Astrolabe* at http://www.uottawa.ca/academic/arts/astrolabe/.

In order to examine this contention we intend to first categorize the use of images in humanities scholarship. We think a few observations about the character of images in the humanities help to reveal some of the difficulties in getting computer-generated methods to receive wide acceptance. Then we will draw out a few broad categories of computer-generated graphics, based on examples from *Literary and Linguistic Computing* and *Computing and the Humanities*, but also a little on our own work.

<Figure 4: Frontispeice of Vico>

Although computer-generated graphics are only a decade or two old, graphical images have played a role in humanities scholarship for centuries. Images that have acted as commentaries on other works (such as the Bible) have been mostly allegorical. Allegories, by their very nature, act as secondary sources to primary materials, in a fashion similar to what computer-generated graphics do to their base texts – they provide another view of the original material, and, if successful, reveal new associations that were contained there. Figure 4 is an example of such an image – the frontispiece to Giambattista Vico's *The New Science*. Vico says in his introduction that the allegorical picture will both help the reader understand the idea of his work before reading it and more easily remember it after having read it. (In other words it will allow the reader to explore the work before reading it and then better remember it after reading it, which sounds a lot like Playfair's comments on the usefulness of graphs for tables of numbers.) Note, that although most of the intellectual material is represented via purely allegorical methods, one of the tools in the allegorical arsenal is space itself – the arrangement of the elements plays a role in defining the associations between them. Metaphysics is *standing* on a globe representing the world of nature. The various objects that represent the civil world are arranged in a line *together* at the bottom of the image. The arrangement in space is a part of the technique used here.

<Figure 5: From Benardete>

We get closer to one of the distinctions we am interested in making when considering schematic graphics, like the drawing from a 20th century philosophical work shown as figure 5 from Seth Benardete's commentary on Plato's *Gorgias* and *Phaedrus*. Clearly, although these figures are significantly more abstract than the Vico, and attempt solely to represent relationships between abstractions, they also use relative positions with simple graphical enhancements in two-dimensional space to represent important associations between different ideas. As long as, in some sense, the relative position of the objects is preserved, minor or even major displacements is immaterial.

In this paper we will call "topological" those graphics which use space principally as room where proximity is used to represent strong associations – following one of the definitions of the word: "concerned with relations between objects abstracted from exact quantitative measurement". Where images are used at all, this type of topological graphic has dominated humanities scholarship for centuries. Furthermore, they have been used as a presentation tools – acting as a visual summary to a written argument. It would appear that their function is pedagogical – to assist a reader/learner to grasp ideas already present in the work. It is, indeed, instructive to compare these examples with George Landow's use of *Intermedia*, where the positioning of hypertextual buttons on the screen is meant to not only allow the student to jump to associated material, but, by the very layout of the buttons on the screen, allow him/her to begin to see associations between the ideas.

In contrast, scientific graphics are much more often "metric" in character – where not only is the relative position of objects, but their exact position in a multi-dimensional space, is important. Their position in the space is exactly specified by a series of numbers – their co-ordinates. Edward Tufte, in his book *The Visual Display of Quantitative Information*, points out that the appearance of the relational graph (where data is represented as points on

an n-dimensional space) is a surprisingly recent development – first appearing in the 18th century – and seems to have arisen from the realization that the plotting of abstract numeric values was similar to the creation of a map. It was not until the 18th century that, even in scientific fields, it became evident that not only spatial and temporal data were suitable for presentation in a spatial manner. Indeed, the first known bar graph, which appears in William Playfair's *Commercial and Political Atlas* of 1786, had to be accompanied by an explanation that stated that none of the dimensions on the graph were "formed by duration."

Survey of Graphical Representations of Texts

Clearly, the metric view of information which characterizes much of the data of interest to the scientific visualization community, is not so easily adapted to the traditional world of humanities scholarship. If one expects to find a community of humanists who have found useful and interesting ways to extract and work with numeric data out of text, you might expect to turn to the ALLC and ACH, and their respective journals – computers are, after all, principally tools that work with numbers. Hence, in order to gain a larger grasp of the range of textually-oriented graphics being produced by humanities research, we examined the Journals of *Literary and Linguistic Computing* and *Computers in the Humanities*.

Many were metric along one dimension. Here are two examples, pulled more or less at random from many:

<Figure 6: Teresa Snelgrove>

In the first, shown in figure 6 and taken from Teresa Snelgrove's article "A Method for the Analysis of the Structure of Narrative Texts"[7], the author, having defined methods to identify narrative modes, has used these techniques to plot the number of occurrences of sustained action vs. general circumstance in George Eliot's *Middlemarch*.

<Figure 7: Nancy Ide>

The second (figure 7), from Nancy Ide's computer-assisted analysis of Blake's *Four Zoas*, shows the pattern of image density and variety in the poem. The solid vertical lines represent the nine Nights in the poems – important structural divisions which she uses in the accompanying discussion of this figure.

Note that both these examples are based on the one dimension which is reasonably close to "continuous" and "metric": time – obstreperous and subjective as it is in a literary work. One could at least, make the argument that the flow of words in a narrative is associated with the reader's subjective sense of time. The "metric" character of time *in* a literary work is, of course, slightly questionable – time *in* a narrative is rarely measured by something as metrical as the ticking of a clock. We tend to think of time in a text topologically, characterized more by events that happen in the sequence of the narrative. Narrative time is, thus, a synthesis of both a metric and topological view simultaneously – events happen close together, or further apart, but their exact clock (metric) time is not as important as their sense of position within events and divisions in the text.

Thus, although narrative-based graphs are relatively common in articles in *CHUM* and the *LLC*, the narrative dimension is often divided into pieces, representing some major structural division in the work (or works) in question – it is simply easier, and probably more important, to relate points in a graph to their position within a work by their position within prominent structural divisions.

Graphics that associate events and structure directly with the narrative itself are likely to be more immediately understandable and convincing than any others. Furthermore, because

---

[7] *Literary and Linguistic Computing*, Vol. 5, No. 3.

they are both understandable rapidly by the viewer, and can be generated rapidly by appropriately designed computer software, they readily support, in turn, an exploratory use of image.

Figure 8 is a graph from our own work presenting the distribution of the various forms of the word "scepticism" through David Hume's *Dialogues Concerning Natural Religion.* It shows the simultaneously metrical and topological nature of narrative time, and the important role that both metric and topological interpretations of the narrative time play.

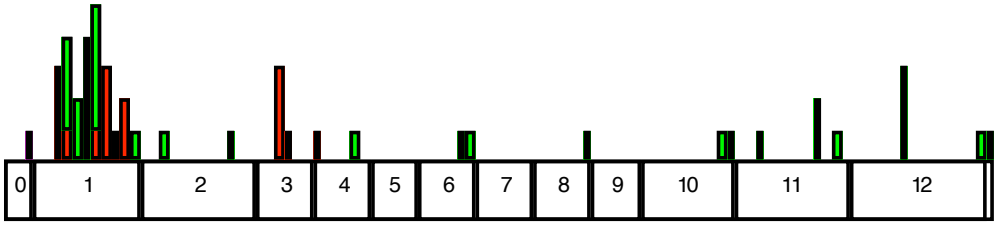

Figure 8: Distribution of "sceptic" in Hume's *Dialogues Concerning Natural Religion*

On this graph:

The height of the column represents the number of uses of a scepticism words on a single page. By plotting based on page number, we could also show concentrations – a column for a page containing three occurrences of one of the "sceptic" words would be three times as high as that for a page containing only one occurrence. On the other hand, it is true that concentrations that spanned pages would be under-represented by this method. There are statistically-based methods that can avoid this problem.

- The colours (shown as different shadings in the printed diagram) represent the speakers: Purple for Pamphilus, Red for Cleanthes, Green for Philo.
- The *Dialogue* is divided into an introduction and 12 parts. The divisions at the bottom represent these sections.

Although the graph can be generated rapidly, and, therefore meets one of the criteria necessary for exploratory use, we were able to observe several things that had not been, to us, observable before:

- Demea does not use the word at all.
- The concentration of the word in part one is probably of no surprise to anyone who knows the *Dialogues* at all, but it may be a surprise that there is no other such concentration in the rest of the text.
- At the end of the work, only Philo uses the word – although, of course, the ending sections are dominated by Philo.

- In sections 4, 6, 8, 10, 11 and 12 the word is used in the "summing up" portions of the section. This is, of course, one of the reasons that Hume can keep the issue of scepticism alive to the reader even through the subject is only relatively infrequently mentioned.

<Figure 9: Thury>

Another prominent source of numeric grist for the graphical mill is frequencies of occurrences of a word within structural entities in the text. Joel Goldfield in his article "Computational Thematics, a Selective Database, & Literary Criticism" (in Rosanne Potter's book *Literary Computing and Literary Criticism*) uses the Z-score as a statistical tool to establish a measure of the how much the number of occurrences of a word or word group in a single textual unit stands out from what might happen purely by chance – and he points out that other scholars such as Alphonse Juilland, Charles Muller and Etienne Brunet had done the same. Thury, in her study of words related to Youth and Old Age in Euripides, uses the same measure – there called the Guiraudian score. She uses graphics as a tool to assist in a vivid comparison between different entities – in figure 9, between "old" and "young" words in the plays. The graphic contains, in some sense, no more data than a simple table would contain – but by presenting the results graphically the reader's ability to compare and balance the numeric data between the different plays is much enhanced. In articles we reviewed comparative graphs are frequently used to facilitate this type of data comparison.

A narrative-based graph or frequency-comparison graphs are both, immediately, more approachable than graphics based entirely on other dimensions. Nonetheless, statistical multivariate analysis (MVA) has traditionally generated a rich collection of 2- or 3-dimensional graphs where all the dimensions were metric.

<Figure 10: Brunet>

Etienne Brunet shows in figure 10 the result of a factor analysis on data based on counts of the number of different types of punctuation for different French authors. The layout of punctuation and author in the same space (the two dimensions of the original data) is characteristic of certain types of MVA analysis, and in this case, by associating punctuation with authors of certain periods, shows how systems of punctuation have changed over two centuries of literature. As a result of the very nature of MVA, it is difficult to associate useful *meanings* to the resultant dimensions – indeed, Etienne Brunet, in his article "What do Statistics Tell Us?" associated with figure 10 acknowledges as much when he says that "he will not put the reader off" by describing the "infinite calculations" necessary to obtain the distances in his factor analysis. As far as we know, only Alastair McKinnon has attempted to associate meanings to the dimensions in his recent MVA analyses of the works of Kierkegaard. By attaching meanings to the dimensions, McKinnon has attempted to define a "conceptual space" for Kierkegaard's ideas. In figure 11 we see some of the data that results from a correspondence analysis of occurrence data based on the distribution of occurrence of certain high frequency words in Kierkegaard's *Fear and Trembling* – plotted so that the distribution over three of the resultant dimensions can be seen.

<Figure 11: McKinnon>

Graphics from MVA result in multidimensional metric spaces that are, in fact, closest to that found in Scientific Visualization data. All the dimensions are metric, and, it is sometimes true that more than two dimensions should be accounted for in interpreting the data. When as few as 3 dimensions come into play, visualizing the positions of various objects in this space becomes difficult, and the ability to, with the computer, turn the space to examine the arrangements of the points is one of a number of appropriate actions. Here is a natural place for 3-D tools of the kind found in Scientific Data visualization.

Preprint of English version published in French as Rockwell, Geoffrey and John Bradley, "Empreintes dans le sable: Visualisation scientifique et analyse de texte", in *Litterature, informatique, lecture* edited by Vuillemin and LeNoble, Paris: Pulim, p. 130-160, 1999. French reprinted online at *L'Astrolabe* at http://www.uottawa.ca/academic/arts/astrolabe/.

Visualization and Text Analysis

As we stated at the beginning it is our contention that the time is ripe for the application of visualization techniques to the analysis of text. We stated that visualization to be accepted, depends on the standardization of data structures and the acceptance of a body of methods, and, as we have shown in the survey above, there seems to be such standardization and acceptance. Narrative-based graphs, frequency-comparison graphs and, to a lesser extent, multidimensional metric spaces, are sufficiently common to suggest that there is the common interpretative ground needed for visualization. Graphical images based on the narrative dimension, images based on the frequency of occurrence of textual objects, and images based on MVA analysis are common in the computer criticism literature we examined. All three types of images could be the basis of "textual visualization" software that could share some of the characteristics of its Scientific Visualization brethren. Furthermore, our examination of graphics that are narrative based suggests that they are well enough grounded in traditional methods that they might, if they were relatively easily produced, meet with wide acceptance and ready use by many of the next ring of humanities computer users – those who use a machine for word processing, but not yet anything else.

As it was for Scientific Visualization, however, the *packaging* of textual visualization tools is very important if this is to happen. It is the absence of textual visualization tools that presently holds back the use of visualizations in the exploration of texts. Few of the text analysis tools that exist have any graphing capabilities. TACT has a rudimentary distribution graph where the bars are made up of asterisks (*); for most visualizations one has to export data from a text analysis package to a spreadsheet or a graphics package. The images are therefore not interactive - they are produced as a result of text analysis and do not assist with the understanding of the text while it is happening. When graphs are produced at the end of the process by exporting data one can hardly use them to explore data interactively; such graphs serve more for communicating results to others than understanding results in the first place. What is needed are tools where the visualizations are generated interactively and, in turn, can be used to trigger other interactions with a text. Imagine if one of the graphs surveyed above were interactive so that clicking on a word would launch a KWIC display where all the instances of the word in the text were listed, or the full text itself. This would be a textual visualization.

Based on the features common in the Scientific Visualization tools that we listed above, we can thus identify the features we should expect in a textual visualization environment:

- **Interoperable Data Exchange**: A collection of standard data structures that represent the basic items of discourse within the system. The TEI SGML implementation could be a starting point for developing such standards.
- **Data Processing Libraries**: A collection of standard operators that work with text. These would be modules that do basic operations like finding patterns, sorting words, and so on.
- **Renderers**: Software modules that take results and display them on the screen in ways that are appropriate to textual visualization. We would expect at a minimum to have modules that could create the types of common graphs surveyed above.

- **A Visual Programming Environment**: A framework environment where users would place and connect the data processing modules and renderers. Users would create the actual explorations in this environment combining modules in a graphical fashion.[8]
- **User Interface Development System**: A programmer's assistant – a collection of routines that can be called by any programmer who is creating his/her own operators or graphic displays. This would allow people to extend the framework by adding to the libraries of data processing and rendering modules.

It is, perhaps, important to realize that the principal user of Scientific Visualization software like AVS is the engineer – not the scientist. She or he is trying to apply general principles of his/her science to a particular problem or object. The research that developed these basic principals upon which AVS is built was done elsewhere. Textual Visualization tools will, like Scientific Visualization, give users access to methods and conventions developed by someone else. Those who are able to break entirely new ground in humanities scholarship, may not find these tools easily applicable. Perhaps this is the reason why we have observed that there are significant pedagogical uses for TACT – it, after all, captures, and then imposes a certain methodology and approach that can often suit a pedagogical purpose, but may not always meet a research one. We would expect that textual visualization software might end up serving the same purpose. It may not be the researchers at the cutting edge who use it, but those who want to quickly see if methods pioneered by others yield interesting resultes on their texts.

Visualization and Reading

In this last section of this chapter we will conclude by looking generally at visualization and reading. We must step back and ask about the theoretical foundations of textual visualization. Assuming that we know what reading is, we must first ask again what visualization is? For our purposes *a visualization is an interactive graphical and metrical representation of data that can be used for exploration.* A textual visualization would thus be a visualization where the object visualized is a text. We can understand this better if we go through the parts of the definition:

- **Interactive Exploration**: First, we expect visualizations to be interactive so that they can be altered and they can trigger other representations, be they textual or graphical. The interactive nature of visualizations is tied to their exploratory use. One studies a visualization by playing with it rather than reading it repeatedly. The practice of interpreting a visualization is one of changing parameters and seeing what changes do to the graph. While there may be one particular view of the data that one settles on as communicating the most, it is the interactivity that both makes a visualization compelling and helps one understand the data.

  Thus the "reading" of visualizations tends to be a much more active practice, akin to exploring a new space. One "looks around", turning multidimensional data rendered on a 2D screen until one has a sense of the whole. In fact, part of the rhetorical effectiveness of a visualization is the perception that the image is not a subjective interpretation of the text, as an illustration is, but that it is metrical - based on measurements of the text. Thus there is the illusion that what one is seeing is the text itself, not someones interpretation of it. This is reinforced by the activity of manipulating the image - one has the feeling of reaching through to the original.

---

[8] We have prototyped what such a framework would look like. See URL: http://www.humanities.mcmaster.ca/~grockwel/ictpaper/ictintro.htm

Preprint of English version published in French as Rockwell, Geoffrey and John Bradley, "Empreintes dans le sable: Visualisation scientifique et analyse de texte", in *Litterature, informatique, lecture* edited by Vuillemin and LeNoble, Paris: Pulim, p. 130-160, 1999. French reprinted online at *L'Astrolabe* at http://www.uottawa.ca/academic/arts/astrolabe/.

- **Graphical Representation**: A visualization is a representation of a text as an image. For the sake of comparison, we could say it is related to textual representations like abstracts and summaries that try to condense some aspect of the full text. A visualization is simply a graphical summary rather than a verbal one.

  Being graphical, a visualization is "read" much the way any diagram or schematization is. One does not read it sequentially starting with one word and proceeding through to the end. Instead images tend to be viewed by moving from the large to the small and back. Edward Tufte in his superb book *The Visual Display of Quantitative Information* writes that excellent graphical displays should, among other things, "encourage the eye to compare different pieces of data ... (and) reveal the data at several levels of detail, from a broad overview to the fine structure"[9]. When viewing a graphic one moves from the overview to the detail and back out. One steps back to gaze at the whole and then one steps forward to look at the place of details in that whole. Graphics can thus be very effective in showing the overall structure of a text and the place of certain details in that structure. They show overall patterns and anomalous details which can provoke rereadings of the text.

  As was noted above, the weakness of graphical representations is that their interpretation can vary widely when there are no standards for the interpretation of secondary features. This is another reason why visualizations are more effective when used by the researcher exploring data than when used by the others out of context. The researcher who created the visualization knows what the graphical features mean. He or she knows what proximity or different colours mean in the context of a particular visualization. By contrast, someone who did not build the visualization usually needs a verbal description as a supplement to set the context for a meaningful reading. It is only when one understands the process by which the image was produced that one can read it with understanding. Even the most common types of visualizations need verbal labels and keys to explain what is being seen. One is tempted to suggest that visualizations are a type of esoteric representation that need a key to be interpreted. The reader once initiated into the interpretative community then has a new and powerful way to understand the text.
- **Metrical Representation**: A visualization is, in order to be interactive and exploratory, usually based on a quantification of the text. The text is measured and it is the measurements that are graphed - the number of occurrences of a particular pattern for example. Hypertexts can be non-metrical interactive graphics where one clicks on parts of an illustration and that triggers an interaction with the system. As mentioned above, it is the metrical character of visualizations that contributes to their rhetorical value for exploration. First, one tends to use them when one has quantitative data in the first place, but also, there is a feeling of immediacy to an image that is based not on someones interpretation, but on a direct measurement of the text. If one disagrees with the visualization one can trace the process back and identify what should be changed in a way one cannot when dealing with an illustration. A visualization environment where the programming is graphically displayed allows one to not only view the results, but also to view the logic of the representative process. One can change the operators or data source if one believes the visualization was improperly generated.

Finally, allow us to raise a fundamental concern that underlies the entire basis of the application of "textual visualization" to the study of literature. In the theoretical part of Jacque Bertin's monumental work *Semiology of Graphics*, he distinguishes between

---

[9] Tufte, *The Visual Display of Quantitative Information*, p. 14.

language, which he calls polysemic, and data graphics, which are monosemic. By "monosemic" he means that each sign in a graphic has only one meaning, and that this meaning is known *prior to* the observation of the graphic. The perceptual process then focuses on the relationships among the signs, and with the outside world. The interpretation takes place entirely among the given meanings, producing a maximum reduction of confusion. As Bertin says, all viewers of the graphic, once they correctly understand the signs, will agree with the meanings of them, and "agree to discuss them no further." In contrast to this, scholarship in the Humanities has been word centered, and word-expressed. Language systems are polysemic – made up of components in the words that cannot, by their very nature, be resolved to a single meaning. Thus, many issues in the humanities are always ready to be interpreted anew – viewed in the light of current culture and understanding of the meaning of the words. Ultimately this fundamental dichotomy between the mathematical/graphical and textual/metaphorical nature of discourse is one with which people who wish to apply scientific and mathematical methods to texts will have to deal.

Preprint of English version published in French as Rockwell, Geoffrey and John Bradley, "Empreintes dans le sable: Visualisation scientifique et analyse de texte", in *Litterature, informatique, lecture* edited by Vuillemin and LeNoble, Paris: Pulim, p. 130-160, 1999. French reprinted online at *L'Astrolabe* at http://www.uottawa.ca/academic/arts/astrolabe/.

Bibliography:

Arnheim, R., *Visual Thinking*, University of California Press: Berkeley and Los Angeles, California, 1969.

Benardete, S., *The Rhetoric of Morality and Philosophy*, University of Chicago Press: Chicago, 1991.

Bertin, J. (Berg, W.J., trans), *Semiology of Graphics,* University of Wisconsin Press: Madison, Wisconsin, 1983.

Brainerd, B., "Textual Analysis and Synthesis by Computer", *Abacus*, 4:2, 8-18.

Brunet, E., "What Do Statistics Tell Us?" in Hockey, S., Ide, N., Lancashire, I., eds, *Research in Humanities Computing,* Vol. 1, Clarendon Press: Oxford, (1991) 70-92.

Earnshaw, R.A., Wiseman, N. *An Introductory Guide to Scientific Visualization,* Springer-Verlag: Berlin, 1992.

Ide, N., "Computer-Assisted Analysis of Blake" in Potter, R., ed, *Literary Computing and Literary Criticism: Theoretical and Practical Essays on Theme and Rhetoric*, University of Pennsylvania Press: Philadelphia, 1989.

McKinnon, A., "Mapping the Dimensions of a Literary Corpus", *Literary and Linguistic Computing*, 4:2 (1989) 73-84.

McKinnon, A., "The Multi-Dimensional Concordance: A New Tool for Literary Research", Computers and the Humanities, 27:3 (1993) 165-183.

Petre, M., "Why Looking Isn't Always Seeing: Readership Skills and Graphical Programming", *Communications of the ACM*, 38:6 (1995) 33-44.

Potter, R., "Literary Criticism and Literary Computing: The Difficulties of a Synthesis", *Computers and the Humanities,* 22 (1988) 91-97.

Potter, R., *Literary Computing and Literary Criticism*, University of Pennsylvania Press: Philadelphia, 1989.

Raymond, D. R. (1993), "Visualizing Texts", *Making Sense of Words*: *Proceedings of the Ninth Annual Conference of the UW Centre for the New OED and Text Research*, (Oxford), 19-32.

Preprint of English version published in French as Rockwell, Geoffrey and John Bradley, "Empreintes dans le sable: Visualisation scientifique et analyse de texte", in *Litterature, informatique, lecture* edited by Vuillemin and LeNoble, Paris: Pulim, p. 130-160, 1999. French reprinted online at *L'Astrolabe* at http://www.uottawa.ca/academic/arts/astrolabe/.

Repenning, Alexander, and Tamara Sumner. "Agentsheets: A Medium for Creating Domain-Oriented Visual Languages." IEEE Computer, March 1995: 17-25.

Bradley, J., Rockwell, G., "What Visualization Teaches us about Text Analysis", Paper presented at ALLC/ACH conference in Paris, 1994.

Snelgrove, T. "A Method for the Analysis of the Structure of Narrative Texts", *Literary and Linguistic Computing,* 5:3 (1990).

Thury, Eva M, "A Study of Words Relating to Youth and Old Age in the Plays of Euripides and its Special Implications for Euripides' *Suppliant Women*", *Computers and the Humanities,* 22 (1988) 293-306.

Tufte, E. R., *The Visual Display of Quantitative Information*, Graphics Press: Chesire, Connecticut, 1983.

Tufte, E. R., *Envisioning Information*, Graphics Press: Chesire, Connecticut, 1990.

Vico, G., *La Scienza Nuova*, Rizzoli: Milan 1977.

Preprint of English version published in French as Rockwell, Geoffrey and John Bradley, "Empreintes dans le sable: Visualisation scientifique et analyse de texte", in *Litterature, informatique, lecture* edited by Vuillemin and LeNoble, Paris: Pulim, p. 130-160, 1999. French reprinted online at *L'Astrolabe* at http://www.uottawa.ca/academic/arts/astrolabe/.

Figures:

<Figure 1: Combination of Table on one side and graph on the other>

<Figure 2: Explorer Screen>

<Figure 3: Prograph Screen>

<Figure 4: Frontispeice of Vico>

<Figure 5: From Benardete>

<Figure 6: Teresa Snelgrove>

<Figure 7: Nancy Ide>

Figure 8: Distribution of "sceptic" in Hume's *Dialogues Concerning Natural Religion*

<Figure 9: Thury>

<Figure 10: Brunet>

<Figure 11: McKinnon>