Appendix A
Computer Techniques

1.10. In this Appendix the computational procedures used to process the text and analyze the imagery of *A Portrait of the Artist as a Young Man* are discussed. These procedures may be divided into three groups: general "housekeeping" programs that could be applied to any text, procedures that facilitate the selection of images and the building of tables making individual images readily accessible for examination, and, finally, programs that perform specific analytic functions. Each of these phases of processing will be described below. I shall attempt to define terms pertaining to computational procedures that might be unfamiliar to the general reader. Whenever possible, descriptions of mathematical procedures will be presented intuitively with references given for more rigorous treatments; models developed specifically in this work will be accompanied by formal definitions as well.

1.20. The text used for this study is the definitive edition of James Joyce's *A Portrait of the Artist as a Young Man,* corrected by Chester G. Anderson from the Dublin Holograph and edited by Richard Ellmann. The text was key-

punched at the University of North Carolina Computation Center, conforming as closely as practical to the format of the printed page. Keypunching was done on I.B.M. 029 upper-case machines using the EBCDIC encoding scheme. Only two keypunching conventions were introduced. To help the computer recognize words and punctuation marks, the text was prepared in blank, delimited form: each individual unit to be recognized by the computer was separated by a space. Text words, of course, are routinely separated by blanks; this convention separates punctuation marks by blanks as well. Spacing would look like this: [look like this :]. Second, a set of characters not expected to be found in the text was used to indicate a change in type font. Passages printed in italics were bounded on the left and right by the "less than" and "greater than" symbols (< and >). To facilitate proofreading, the text was punched with one line of printed text per standard eighty-character I.B.M. card. The total text constituted some ten thousand punched cards.

1.21. After the text was prepared in a computer-accessible form, the next step was to separate it into distinct entities (*i.e.*, words or punctuation marks) and to index them.[1] The text was introduced into the machine one card at a time, and the computer then scanned the card for a set of characters bounded by blanks. Each word or punctuation mark located by this procedure was placed in a "logical record" along with a number indicating where in the text it occurred; this record was then written onto magnetic tape[2] for additional processing and a printed version was made for manual reference. The particular format used was as follows:

2	6	3	7	18
Length of Word	Linear Sequence Number	Page Number	Blank	Word

A two-digit number indicating the length of the word was followed by a six-digit number indicating the relative sequence of the word or punctuation mark in the text. (The first word is numbered 1, the second 2, etc. until the end of the text.) The page number of the printed text on which the word appeared was included for manual reference. Finally, the word or punctuation mark was stored in an eighteen-character slot. (No word appearing in the text was longer.) As indexed by this scheme there were slightly more than 98,000 words and punctuation marks in the text.

1.22. Next, the records were sorted primarily on word and secondarily on linear sequence number, using the standard system/360 sort package furnished by I.B.M. The sorted records would thus be in alphabetical order with repetitions of the same word-type appearing in text order. Again these records were stored on tape.

1.23. The final program of the text preparation phase was a modified version of Sally Sedelow's SUFFIX program. This procedure groups words together by root or stem and discards all function words and punctuation marks. Each group of words with the same root but with different suffixes is identified by a five-digit number, called a MATCH-COUNT, which is attached to the logical record for each word. Thus each appearance of *complete, completely, completing,* etc. would be recognized as belonging to the same root group by the shared MATCHCOUNT. Also attached was a number indicating the total frequency for the particular word-type. Output records, again stored on tape, had the following format:

Length of Word	Linear Sequence Number	Page Number	Match Count	Frequency	Word
2	6	3	5	4	18

One record was created for each occurrence of each word; however, at this point function words and punctuation marks were discarded from the data set. This program marks the end of the basic data preparation steps.

1.30. The programs of phase two were designed to make individual images and the images around them readily available for examination. The first step was the selection of the words to be considered images. This was done manually and represents the only major task performed manually in the analysis, except, of course, the initial keypunching. It would have been desirable to have given the computer a definition of *image*, a text, and perhaps a dictionary, and to have allowed it to select the images using these criteria. Unfortunately, the state of the art does not make this approach feasible at the present time; recent work on thesauri and dictionaries, however, indicates that this may eventually be possible.

From an alphabetical listing those words were selected which I felt had sensory and thematic import. No systematic restrictions were placed on images: for example, no distinctions were made among *flame, flames, flamed, flaming, etc.* all were considered as variant forms of the single image *flame*. Several other persons have examined my selections; their fruitful suggestions often resulted in reconsiderations of individual words. Inevitably, some will disagree with my selections, but I have attempted to err on the side of over-inclusion rather than risk missing an important facet of the novel by leaving out some particular image. The final result is a list of some 1,312 images that is, I feel, a reasonable basis for the analysis. (A complete list of images is provided in Appendix B.)

1.32. The computer was furnished with a list of MATCHCOUNTS for those words considered to be images.

The list consisted of a series of cards, one image—or MATCHCOUNT—per card. To facilitate this process and to avoid keypunching 1,312 cards, I had the computer punch a card for each word-type in the text with its MATCHCOUNT. I then selected the cards for the image words and discarded the rest. These cards were read in ascending order of MATCHCOUNT; the data set from the suffix program was then passed against this list. Records with MATCHCOUNTS that also appeared on the list of image MATCHCOUNTS were selected and stored in a dictionary of images called IMG:

```
01 IMG
   02 MATCHCOUNT
   02 IMAGE
   02 FREQUENCY
   02 BEGIN
   02 FINISH
```

The index information for each occurrence of the image was separated and stored in a long list called LOC1. The starting position in this list of the index information (the linear number, page number, etc.) was placed in BEGIN, and the position of the last slot in the list where index information for an occurrence of the image group was placed, in FINISH. The process was repeated for all 1,312 images. Using the two lists, I could locate the textual information for all occurrences of any image by going to the slots in LOC1 between the particular BEGIN and FINISH numbers for that particular image. This organization is shown in Figure A.1 for the image *abyss*.

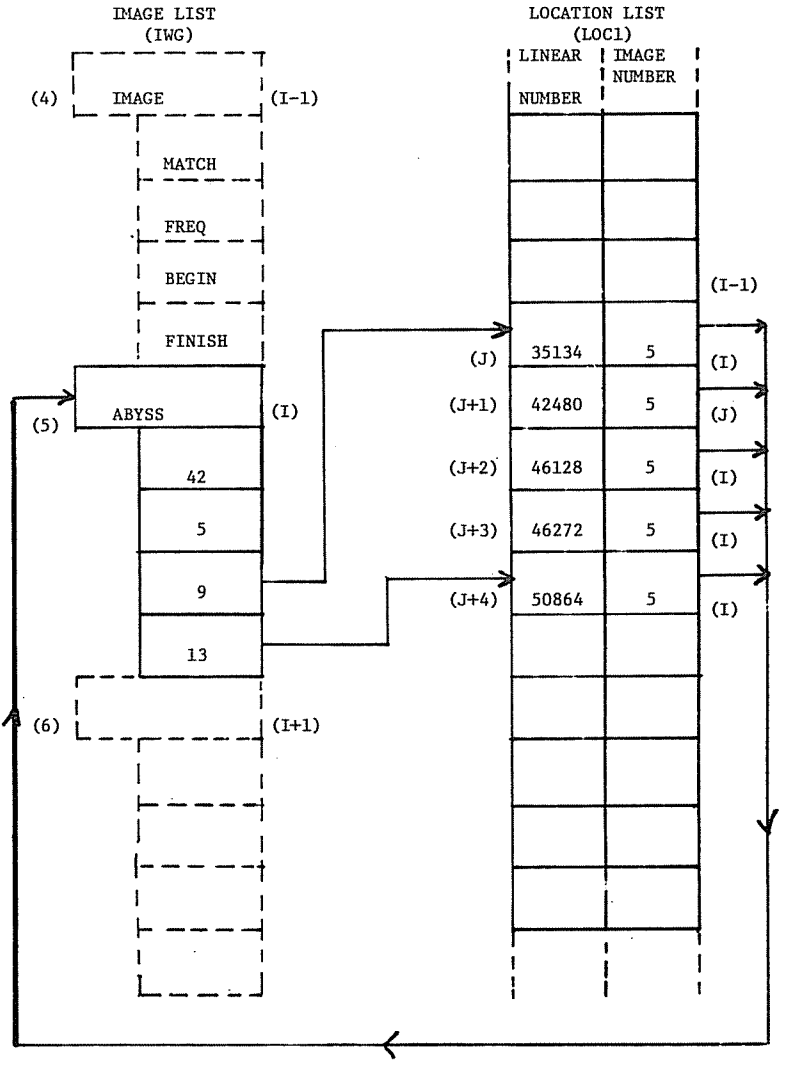


Figure A.1 Image and Location Tables

1.33. The list, LOC1, is in image order. A duplicate list of index information was created and sorted into linear text order; that is, the first entry would be the index information for the first image; the second entry, the information for the second image, etc. Using parts of all three lists I could now examine the environment of any image. Thus, I could look at, say, the five images before and after each occurrence of the image *fire*. To do this easily, however, I needed a number or pointer connecting the slot in LOC1 with the position in LOC2 for that particular image. When this number is attached, all three lists are connected logically by the numbers or pointers; the complete structure is shown in Figure A.2.

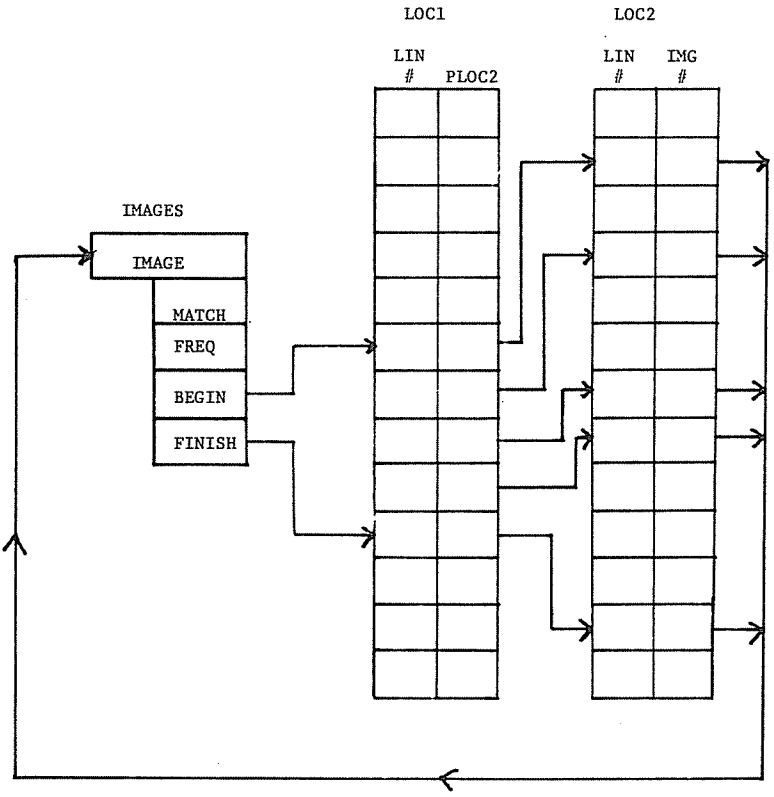


Figure A.2. Complete Image Tables

1.34. These techniques, although written specifically for this analysis, could be applied to any study where one wished to store and examine a list of entities and their contexts. A generalized version of these programs has been written and described in *Computers and the Humanities*.[3] The completion of these steps marked the end of the "housekeeping" data processing chores; a discussion of programs used to analyze *Portrait* follows.

1.40. The first set of programs written specifically for this analysis produced image-frequency counts for each chapter of the novel. These frequency counts help determine for individual images broad patterns of distribution. If several images appear a large number of times in some chapters but not in others, we would suspect some sort of associative relation among them; such leads were pursued more thoroughly using analytic techniques to be discussed. (See Appendix B for a list of images and their frequencies of occurrence by chapter; the frequency listed is the cumulative frequency for all variant forms of images having the same stem indicated by similar MATCHCOUNTS.) The various counts were produced by checking each image position in LOC1 to see in which chapter that particular linear number fell, and counters for the five chapters were incremented accordingly. The computer produced a printed record of this information as well as a punched-card record, one image-type per card with its accompanying frequencies. These records were then sorted five different times, each time on the decreasing frequency of occurrence for a particular chapter, to produce lists of images in descending frequency order for each chapter. Helpful in themselves, these listings facilitated the selection of thematic groups of images used in subsequent programs.

1.41. The second major analytic program developed

groups or clusters of images that consistently occur close to one another within individual chapters. A rather elaborate model called factor analysis or, more precisely, principal component analysis, was used. If a group of images consistently appear within a hundred words of one another in a chapter, this would suggest that they are related in Stephen's experience. The principal component program reveals such tendencies. (See Appendix D for a list of factors developed.) The program was executed, with slightly varying lists of images, for each of *Portrait*'s five chapters.[4] The major programming for this step was a procedure that computed the data used as input for the "canned" factor analysis procedure.

These programs were run for each chapter. Using the image frequency lists described above, I selected the individual images to be analyzed; the images chosen for each chapter were those that occurred more than a specified number of times in that chapter. (This number ranged from eight to five occurrences per chapter for the five different runs.) By this process some 90 to 120 images were examined for clustering tendencies in each chapter.

The text was divided into 100 word segments (words 1–100, 101–200, etc.). Using IMG and LOC2, the program then determined the frequency of occurrence for each selected image for each textual unit of the chapter. A set of counters, one for each image being considered, was established; the linear numbers and dictionary pointers in LOC2 were considered sequentially. So long as the linear number for the image was within the first 100 words of the novel, the appropriate counter was incremented by one. When the first linear number considered became greater than 100, a record of all the counter values was stored in auxiliary storage and the counters reset to zero. After all subsections of a

chapter were thus examined, the data were then passed as input in the form of a list of numbers, or matrix, to the factor analysis program. Each row of the list or matrix represented the number of times that individual images appeared in a particular subsection of the text. There would be as many rows of numbers as subsections of text, and as many columns as images chosen for examination. Thus if one were interested in M different images and divided the text into N subsections, the list or matrix would be dimensioned MxN with MxN individual cells in it. See Figure A.3.

The principal component program itself looks at each pair of images in all text subsections and assigns the pair a value ranging from -1 to $+1$ (this value is called a correlation coefficient).[5] If the images consistently occur together in the same context, the correlation coefficient is near $+1$; if they never occur in the same environment, the correlation coefficient is near -1; a random distribution results in a correlation coefficient near 0. This process, then, reduces the NxM to a square (MxM) matrix called the correlation matrix. See Figure A.4.

	image 1	image 2	image 3	.....	image M
section 1	f 11	f 12	f 13	.....	f 1m
section 2	f 21	f 22	f 23	.....	f 2m
section 3	f 31	f 32	f 33	.....	f 3m
.	.	.	.		.
.	.	.	.		.
section N	f n1	f n2	f n3	.....	f nm

Figure A.3

Subsequent steps in the process are probably easiest understood in terms of their geometric or vector analogue. We may regard each row of the correlation matrix as an ordered set of numbers (a_{11}, a_{12}, . . . , a_{1m}), or as a point in a Euclidean space of dimension M, or as a vector. If one regards each row as a vector, then the set of all M vectors (one for each row) will generate a space of dimension D, such that D L M. For example,

	image 1	image 2	image 3	. . . . .	image m
image 1	a 11	a 12	a 13	. . . . .	a 1m
image 2	a 21	a 22	a 23	. . . . .	a 2m
image 3	a 31	a 32	a 33	. . . . .	a 3m
.	.	.	.		.
.	.	.	.		.
image m	a m1	a m2	a m3	. . . . .	a mm

Figure A.4

the three vectors given in Figure A.5 could be said to generate the usual three-dimensional Euclidean space since any point or any vector in the space could be generated by taking combinations of the three vectors

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \begin{matrix} a_1 \\ a_2 \\ a_3 \end{matrix}$$

Figure A.5

given. For example b = (2,2,3) can be represented by $2a_1 + a_2 + a_3 = 2(1,0,0) + (0,2,0) + (0,0,3) = (2,2,3)$. (See Figure A.6.) In general M vectors will generate a space of dimensionality less than or equal to M.

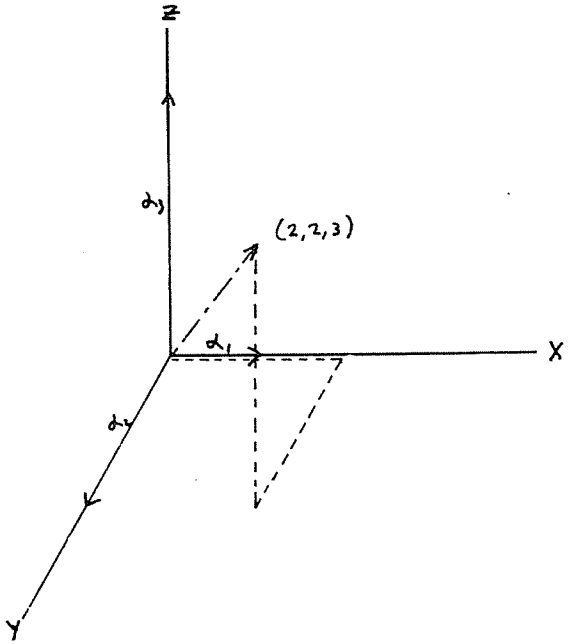


Figure A.6 Linear Sum of the Vectors

The factor-analysis model seeks a group of vectors, formed by various combinations of the original vectors, that comes nearest to generating the original space of the correlation matrix. This approximation is close when a number of the original vectors lie relatively near to one another. In 2-space, this process might be represented as shown in Figure A.7 where the original vectors, $(a_1, a_2 \ldots, a_8)$ might be approximated or reduced to β_1, β_2, with a_5 largely excluded.

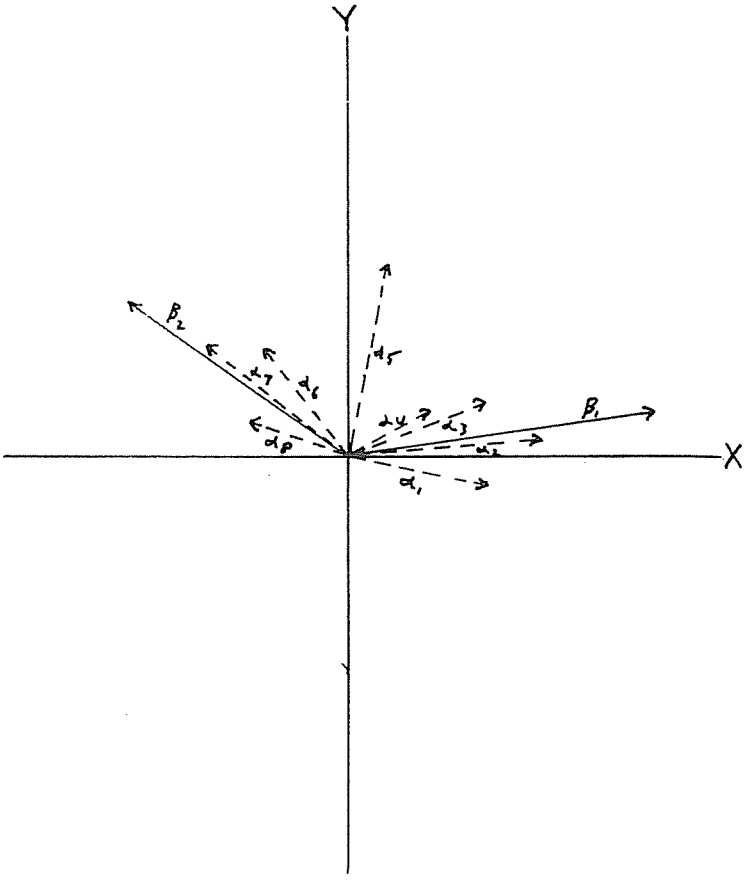


Figure A.7. Resolution of Vectors

What the program actually produces is a set of column vectors (or factor loadings) of the form shown in Figure A.8. Each element or weight of the factor represents the degree to which that particular variable (image, in our case) contributes to the factor. Thus individual factors can be thought to be most strongly characterized by those variables or images which contribute the largest weights. A negative weight implies the absence of that variable (or image) in the context of the variables

Figure A.8

word	a
1	11
word	a
2	21
word	a
3	31
.	.
.	.
.	.
word	a
n	n1

Figure A.9

word	
1	.05
2	.91
3	.63
4	.81
5	.53
6	.66
7	.32
8	.21
9	—.55
10	.11

or images that most strongly characterize the factor. See Figure A.9: this factor is best defined in conjunction with images 2, 4, 6, 3, 9, and 5: however, image 9 consistently *does not* appear in context with the others.

The set of factors developed for each chapter represents clusters of images that consistently appeared together. These groupings were verified by using an image concordance. Thus, by using the frequency lists, the set of factors indicating tendencies of images to occur close to one another consistently over a chapter, and a concordance listing each occurrence of each image with the five images on each side, I was able to infer the changing patterns of associations among images that develop over the novel and to verify these patterns by going back to the textual context for each occurrence.

1.42. The concordance program was virtually a trivial programming task once the data were put into the random accessible structure mentioned. This was done by having the computer look at each image, sequentially, in the dictionary. From there, it looked at each occurrence of each image in LOC1. From there, using the pointer to LOC2, it located the textual position of each occurrence, moved out five positions on each side of the particular image in LOC2, and then printed the sequence of images found. This process was repeated for each occurrence of each image.

1.43. To show that important moments in the development of Stephen's personality are accompanied by large concentrations of important images, a model was developed to quantify "richness" of imagery. Richness, in this context, is a function of both the total number of images present in a section of text as well as the relative "importance" of the individual images themselves. Since this model was de-

veloped for this project, it will be defined formally; a brief discussion of the computer implementation follows.

A grid was imposed over the text dividing it into units of 500 words, thus giving a resolution of just a little over a page of text per unit. The intent of the model was to represent the entire set of images as a space—or more precisely, as a geometric solid, embedded in a space, for which the volume could be computed. Consequently, the relative volumes of the images in a subset of the text could be computed and would indicate proportional richness of imagery. This was done in the following manner:

1. Let each image be represented as a point in some space such that each unique image adds a dimension to that space. Thus for N unique images, the space would be of dimensionality N.

2. Let the vectors associated with each point or image be orthogonal to one another and of the form (O, O, . . ., a_i, O, . . . , O).

This configuration can be thought of as a mathematical extension of the familiar three-dimensional Euclidean space, where the vectors associated with each of the images would lie along one of the axes (see Figure A.10). Thus for N images, the space would be defined by N such axes and would be of dimension N.

3. The space generated will be such that each point in it will be represented by an ordered N-tuple (a_1,a_2, a_3, . . . , a_n)).

4. This space is representative of the relative frequencies of the images of a_i = frequency of image (i) for some subset of the total text. Thus $O \leq a_i \leq f_i$ where f_i is the total text frequency of image (i).

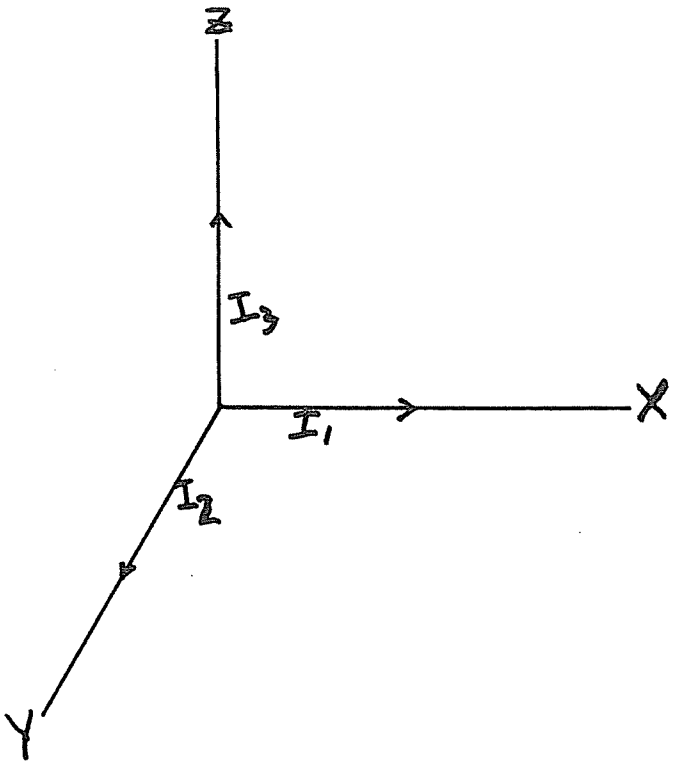


Figure A.10. Vector Representation of Images

Each point in this space, then, represents some collection of images. In our 3-space example, the point, (1, 0, 2), may be associated with some section of the text in which Image (1) appears once, image (2) not at all, and image (3) twice

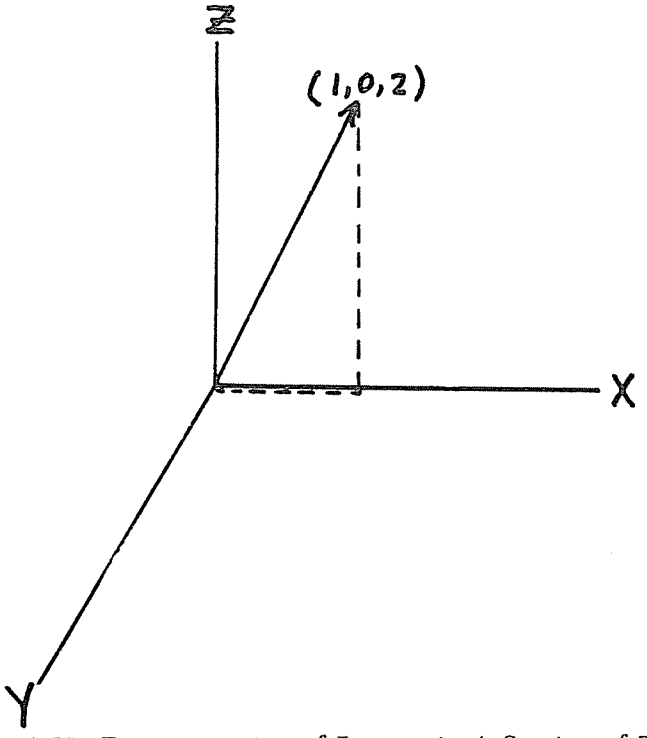


Figure A.11. Representation of Images in A Section of Text

(see Figure A.11). Thus we can associate each of the subsections of the text with a point in the relative frequency space defined by an ordered list of the frequencies of the 1,312 images of *Portrait* for that text subsection.

Not all images are equally "important" in the novel. Some that appear only one or two times carry relatively little weight as compared with images such as fire and water.

which are found throughout the novel in a variety of key passages. To make the model more sensitive to these images, a weighting scheme was imposed to make passages that contain a high concentration of frequently used images stand out from passages that might contain an equal number of images but images that merely represent a cross section of the total set of images.

5. Images were weighed using a function of their total text frequency This function may be represented as a mapping onto a weight-frequency space. Such a function is :

$$v\,(a_1\ \ a_2\ \ \ldots\ \ a_n) = (f_1^{\,a_1}\ \ \ f_2^{\,a_2},\ \ldots,\ f_n^{\,a_n})$$

This space has the additional feature that it is defined multiplicatively rather than additively, as was the relative frequency space, so that deviations from the norm are indicated more dramatically.

6. The "volume" associated with the set of images or any subset of them can be computed. For a given subset, defined by the point:

$$\alpha_1 = (a_{11}\ \ a_{12}\ \ \ldots\ \ a_{1n}),$$
$$v\,(\alpha_1) = (f_1^{\,a_{11}}\ f_2^{\,a_{12}}\ ,\ \ldots\ \ f_n^{\,a_{n1}})$$

This point can be projected onto the corresponding unit vectors and the volume of the parallelapiped formed computed. The volume for the first textual subset would be:

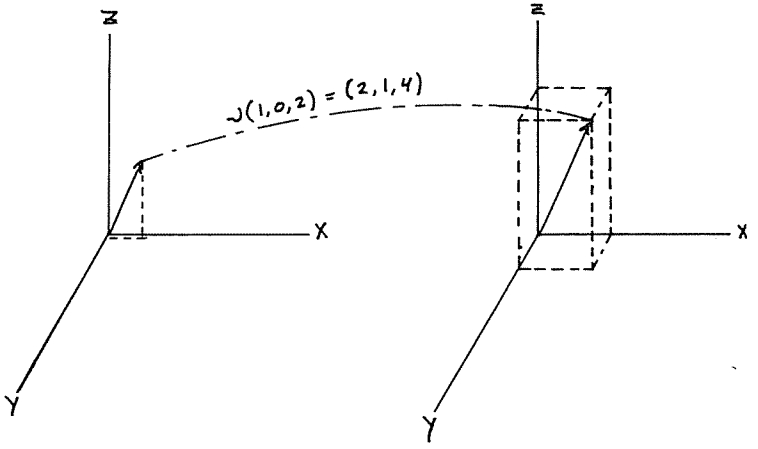
$$B_1 = \begin{pmatrix} a_{11} & & & & \\ f_1 & 0 & 0 & \dots & 0 \\ & a_{12} & 0 & \dots & 0 \\ 0 & f_2 & & & \\ \vdots & & & & a_{1n} \\ 0 & 0 & 0 & \dots & f_n \end{pmatrix}$$

$$\begin{aligned} \text{VOL}(B_1) &= |B_1 \cdot B_1'| \\ \text{But since } B_1 &= B_1', \\ \text{VOL}(B_1) &= |B_1| \end{aligned}$$

$$= \begin{matrix} a_{11} & a_{12} & a_{13} & & a_{1n} \\ f_1 & f_2 & f_3 & \dots & f_n \end{matrix}$$

Thus, for our familiar example, if each of the images occurred twice in the whole text: $(1,0,2) = (2^1,2^0,2^2) = (2,1,4,)$. The point (2,1,4) is projected onto the axis. The volume of the parallelapiped, defined by the projections onto the axis,

is then computed for each of the subsections of the text (see Figure A.12). This calculation is greatly simplified by using matrix algebra, as demonstrated above.



Figue A.12. The Mapping Function, U

7. By conceiving of a text as a linear sequence of words, we can impose a grid over the text that divides it into M equal subsections.

8. The volume associated with each of the subsections, computed in the B space, will reflect the relative richness of imagery in that subsection, where richness is related to both total number of images within a subsection and the relative "importance" of those images. Based upon these assumptions, the model produces a parameter for textual richness objectively derived within the terms of the model.

The model has several rather nice features. The additive identity, zero, maps onto the multiplicative identity, 1, as would be expected. Also, the model makes interpretation of points outside of the original set possible. For a given text, T, $0 \leqslant a_i \leqslant f_i$. But if we hold the set of weights, F, constant, we may then apply them, through the mapping function, to any text. This would make comparisons for several works relative to some particular work possible. For example, it would be possible to measure the richness of the images of *Portrait* that also appears in *Ulysses.*

Actual omputation is far less complicated than the model. A set of accumulators, initiated at 1, is defined—one for each image in the text. Then a scan of either list of image locations is made. The linear number is divided by the unit (in this case, 500) and the corresponding accumulator, plus one, is multiplied by the total text frequency of the particular image. (For an image with linear number 2037, 2037/500 $=4$, accumulator 5 is multiplied by the frequency of that image.) The numbers developed are very large—in fact, a scaling procedure was used to avoid overflow—so that the value actually plotted on the graph is the logarithm of the number. The results of this procedure applied to *A Portrait of the Artist* can be found in chapter 2.

1.44. Often images with the same denotative meaning carry similar connotative values. For example, *burn, burning, blaze,* and *flame* all function virtually alike in the associative

structure of Stephen's mind. Some twenty-seven such thematic groups of images were defined and their respective collective distributions over the text computed, using units of 500 words per linear subsection.

To facilitate interpretation of the characteristic pattern inherent in the raw data for each theme, a technique known as Fourier analysis was used. Numerous discussions are available that include the equations for this model; a brief intuitive description follows. Any wave form, such as that produced by a voice spectrograph or an oscilloscope, that is made up of a finite number of points can be reproduced *exactly* by a combination of perfectly regular sine and cosine curves of different frequencies and amplitudes. That is, if a wave form is made up of some 1,000 points or observational values, that pattern, no matter how ocmplex or "ragged" it looks, can be duplicated exactly by no more than 1,000 sine and cosine curves with frequencies 1, 2, 3, ..., 1,000 cycles per unit and with approriate amplitudes. Accompanying the computer analysis of any given wave form—the graph of a thematic distribution over a text may be regarded as a complex wave form—are numerical values indicating the amplitude or power associated with each frequency. By picking the eight to twelve largest amplitudes and then plotting the associated sine and cosine curves for the respective frequencies, we can obtain a "smoothed" form of the wave that indicates its characteristic pattern. Appendix E contains graphs of thematic groups of images and their accompanying characteristic curves enabled me to determine patterns of interrelations and associations among not just single images but whole groups of images.

1.45. The lists of images with their frequencies in each chapter (1.41), the factor analytic procedure (1.42), and the concordance (1.43) are all used to examine the environments

in which images occur, under the assumption that if images appear close together a number of times they are related in Stephen's experience. This analysis is done to trace the changing patterns of associations among images in Stephen's mind, thus revealing the structure of his personality. The analysis of richness of images (1.44) is used to test the hypothesis that major transitions in Stephen's development are accompanied by dramatic build-ups of important images. These procedures and the materials they produce are then used as tools by the critic to test and develop the ideas and hunches he has derived from his reading and contemplation of the novel.