Estimating Changes in Collocations of Key Words Across a Large Text: A Case Study of Coleridge's Notebooks

David S. Miall

Department of English, University of Alberta, Edmonton, Alberta, Canada, T6G 2E5
e-mail:dmiall@ualtavm.bitnet

Abstract: Existing methods for text analysis, based on the *z*-score, are used to identify significant collocates of emotion words in the notebooks of Coleridge. The collocates found are shown to be both distinctive to Coleridge's lexicon for emotion and to differ from the lexicon of a sample of representative contemporary writings on emotion. A new method is then described for mapping changes in the collocational environment of a set of key words across the text of the notebooks: shifts in selected collocates are examined and graphed across each year of the notebooks (from 1794 to 1819). The graphs point to changes in Coleridge's psychological understanding of the role of emotion during this period: a development from a phenomenal to a process view, then to an integrated and systematic account of emotion in relation to other human faculties.

Key Words: text analysis, change-point analysis, collocation, emotion, Romanticism, Coleridge, psychology.

Introduction

An examination of the frequency and distribution of a word or group of words across a text is a common starting point for the computer-aided analysis of texts. Finding that a given word plays a significant role within some thematic thread running through the text, the analyst will then be interested in studying the lexical environment of that word. What are its significant collocates, and do the collocates change in some meaningful way as the theme is developed across the text? Techniques for encompassing the last of these objectives, however, are not readily available within existing text analysis software, such as the *OCP*, *WordCruncher*, or *TACT*. In this paper I report a simple but effective method for revealing changes in selected collocations across a large text.

The present report forms part of a larger study of the development of Coleridge's understanding of psychology. Coleridge (1772—1834) is well-known as one of the important sources of modern literary theory, primarily due to his seminal (if unwieldy) study, *Biographia Literaria* (1817). But Coleridge's views of poetry underwent significant changes prior to this work, and were influenced both by his own experience as a poet and by his wide reading in philosophy and literature. In his twenties a follower of Associationist theory, as espoused by the 18th Century writer David Hartley, Coleridge turned in his thirties to German idealist philosophy and radically revised many of his earlier views. The mature literary theory of 1817 is firmly based on an idealist account of the mind. So far, however, little attention has been paid to Coleridge the psychologist. While Coleridge's astuteness as a psychologist has always been acknowledged, no systematic study of the development of Coleridge's thinking as a psychologist has taken place.

An important component of Coleridge's psychological understanding was provided by his own experience. Coleridge frequently examined the processes of his consciousness and probed the motives of himself and others. Such reflections,

David S. Miall is associate professor of English at the University of Alberta. His research interests include reader response, the teaching of English, and Romantic literature, especially the poetry and prose of Coleridge. He is editor of Humanities and the Computer: New Directions *(Oxford University Press, 1990).*

Computers and the Humanities **26**: 1—12, 1992.

coloured with the vividness of immediate experience, are scattered in profusion through the notebooks that he kept throughout his life. The notebooks, so far published in four volumes covering the period 1794 through 1826, provided the main source for this study. Volumes 1 to 3 were made available to me in machine readable form.[1] In preparing for the study I coded the material by date as well as by note number in order to keep track of the period at which a given note was written. The notebooks formed a valuable source for tracing the development of this author's ideas across an extended period of time: a type of study which is common enough using conventional means of research, but which seems absent so far in the domain of computer-aided research.[2]

The specific focus of the study to date has been Coleridge's understanding of emotion. Coleridge saw himself, and was seen by others, as an expert in this domain: "you are to be the historian of the Philosophy of feeling," Humphry Davy wrote to him in 1804.[3] The notebooks contain much evidence of Coleridge's interest in emotion and its relation to other aspects of the mind, and close inquiry suggests that Coleridge's understanding of emotion played a key role in the development of his principal ideas. An important strand in the poetic theory of 1817 is the role of emotion in the process of writing or reading literature. In the formal prose of the *Biographia Literaria*, the emphasis on emotion is sometimes toned down or eliminated, but the original notes on which such passages are based often show the centrality of Coleridge's conceptions about emotion. For example, in Chapter 14 of the *Biographia*, Coleridge offered a definition of poetry based on the formal properties of verse — on meter and on the relation of part to whole. In the first draft of the passage in a notebook of 1809, however, emotion forms his basic premise:

> Poetry is the species of composition, which represents external nature, or the human mind, — both in relation to human affections — so as to produce immediate pleasure — /and the greatest quantity of immediate pleasure in each part, that is compatible with the largest possible ‹sum› of Pleasure in the whole. — (NB.iii.3615)[4]

In rewriting the definition for the *Biographia* Coleridge dropped the phrase "human affections."[5] A review of this and other passages suggests that emotion is central to the development of Coleridge's ideas, and at the heart of his psychological thinking.

The range and scope of the notebooks is extensive. The use of computer-based text retrieval and analysis is thus a useful aid to the study of specific themes. The study to be described consisted of two main phases: first, the use of *z*-scores and other methods to examine the lexical environment surrounding Coleridge's use of emotion words in general in the notebooks. Having identified in this way a range of significant collocates, the second phase involved tracking their distribution across the material year by year. The primary aim was to locate changes in Coleridge's vocabulary for matters of emotion across the period of the notebooks, and then to establish whether these changes were systematic and significant.

Z-scores and Other Preliminaries

The Notebooks in their original form include some lengthy passages in Latin, German, Italian, and other languages besides English. For the purpose of this study, an electronic version of the Notebooks was prepared in which all such passages were deleted, except for minor phrases embedded within English material; the size of the resulting text was 457,065 words.

As a first step, data was extracted on the occurrence of all words collocating with emotion words in the notebooks. The emotion words chosen were: *feel, feels, feeling, feelings, felt, emotion, emotions, passion, passions*, and *passionate* (the words occur a total of 1046 times, the commonest being *feeling* with 308 occurrences). These are the words that Coleridge invariably uses when referring to emotion in general (a range of words for specific emotions also occur, of course, but a detailed examination of these lay outside the scope of this study). The collocation span was set at five words either side of an emotion word and all words were collected that occurred within this span. Words occurring only once were deleted from the list. A *z*-score was then computed for the remaining words, to help assess the significance of each collocation, using the formula of Berry-Rogghe.[6] A total of 231 words had a *z*-score of

2.57 or above (the significance tables for the z distribution show that a word occurring at $z = 2.57$ can be assigned a 0.01 per cent likelihood of collocating by chance). The words were then sorted by z-score.

The resulting list provided a useful sense of Coleridge's vocabulary when referring to emotions in the notebooks. A number of the collocations stood out immediately as embodying typical Coleridgean concerns. He described feelings as "deep" ($z = 7.50$) or "obscure" ($z = 9.05$) when attempting to locate those feelings that play a formative role in our philosophical development. In the evolution of his poetic theory, the "modifying power" of the imagination is mentioned in the *Biographia*. The word "modify" ($z = 9.33$) in the notebooks is used in this sense in collocation with emotion: the imagination, for example, is described as the "power of modifying one image or feeling by the precedent or following ones" (NB.iii.3247). An early preoccupation of Coleridge, as mentioned above, was Associationist theory, and even after Coleridge had emancipated himself from Hartley's theory he continued to recur frequently to association as a problem. Thus "associated" ($z = 7.94$) occurs as a significant collocate of feeling. In this and other respects, the list thus agrees with a number of the known currents of Coleridge's thought.

The list also, however, contains a range of other words whose apparently significant role in relation to emotion has been less often remarked. If Coleridge is developing an account of emotion in the notebooks, the question arises whether such words form a unique part of this strand of Coleridge's thought, or whether they also contribute to his general discussions of the mind and of poetic theory. To check this, collocation lists for several other major groups of words were also compiled: words relating to the mind, to poetry and poets, to language, and to love. These target words were selected on two grounds: first, they occur with sufficient frequency through the notebooks, and second, they play a central role in Coleridge's discussions of psychological and philosophical matters.[7] The z-scores of words occurring in the five collocation lists were compared. The first part of the list is shown in Table 1: this has been edited to show only those items that collocated with emotion words three or more times. The list served to reveal a number of words that do indeed seem to be distinctive to Coleridge's discussions of emotion, such as *bodily, deeply, sensation, distinctly, touch,* and many others.

It could also be argued, however, that Coleridge might be using the same vocabulary for emotion as other writers of his period. To check this, collocation lists of emotion words were also compiled for a group of texts contemporary with the notebooks: two by Wordsworth (a selection of critical prose, including the Preface to the *Lyrical Ballads, and The Prelude*), and a novel, *The Monk*, by Matthew Lewis (this choice of contemporary texts was limited by the texts available to me in electronic form). Each of the texts was expected to overlap to some degree with Coleridge's interests in emotion. As in the previous comparison, however, a range of words was found that were significant to Coleridge's discussions of emotion but were not significant collocates of emotion in the other texts. A similar comparison was also made with collocates of emotion words in Coleridge's complete poetry: interestingly, there were no important overlaps of collocates between the notebooks and poetry either, suggesting that Coleridge's handling of emotional material differs rather markedly between the two styles of discourse.

Among the words occurring as significant collocates of emotion in the notebooks ($z = >2.57$), the following were either absent or non-significant collocates of the target words in the other collocate lists — the Coleridge notebook lists of non-emotion collocates and the collocates of emotion words in other texts (including Coleridge's poetry): *associated, belonging, bodily, consciousness, dimness, distinctly, duty, imagination, mental, mind, modify, pleasurable, vivid*. It follows that these words are, on the basis of this test, among those that are likely to be distinctive to Coleridge's vocabulary for discussing emotion in the notebooks. Each occurred three or more times in significant collocation with an emotion word, as measured by z-score.

Changes in Collocation across the Text

The whole range of collocations was examined in preparation for the second phase of the study, however, not only those occurring in the short list

TABLE 1
Comparison of *z*-scores for selected collocates of emotion words in Coleridge's notebooks, compared with four other groups of words

	emotion		love		mind		poet		language	
	Z	F	Z	F	Z	F	Z	F	Z	F
pleasurable	16.64	13	—	—	—	—	—	—	—	—
vindictive	11.87	4	—	—	—	—	—	—	—	—
bodily	10.34	11	—	—	2.07	2	—	—	—	—
pang	9.64	6	—	—	—	—	—	—	—	—
modify	9.33	5	—	—	—	—	—	—	—	—
obscure	9.05	9	—	—	—	—	—	—	2.97	3
connected	8.37	11	—	—	—	—	—	—	—	—
generate	7.99	3	6.50	2	—	—	—	—	—	—
associated	7.94	5	—	—	—	—	—	—	—	—
myself	7.78	20	2.52	7	—	—	—	—	−0.12	3
deeply	7.55	9	—	—	1.82	2	—	—	1.24	2
deep	7.50	18	1.62	5	1.82	2	—	—	1.24	2
excites	7.04	4	—	—	—	—	4.89	2	—	—
moral	7.00	20	−0.67	2	—	—	4.18	9	−0.36	3
strongly	6.97	5	—	—	—	—	—	—	—	—
sensation	6.86	10	—	—	—	—	—	—	0.75	2
sense	6.85	29	−0.36	5	−0.81	3	−1.18	2	8.14	28
incidents	6.82	3	—	—	—	—	6.42	2	—	—
malignant	6.82	3	—	—	—	—	—	—	—	—
distinctly	6.81	6	—	—	—	—	—	—	—	—
impulse	6.67	6	6.94	5	—	—	—	—	2.79	2
bad	6.44	12	—	—	—	—	3.53	5	—	—
interpreted	6.38	3	—	—	—	—	—	—	7.35	3
remote	6.38	3	—	—	—	—	—	—	—	—
sympathy	6.36	9	1.07	2	1.40	2	2.65	3	—	—
excitement	6.01	4	—	—	—	—	—	—	—	—
tranquillity	6.01	3	—	—	—	—	—	—	—	—
awe	5.80	5	7.38	5	—	—	—	—	—	—
consequent	5.80	5	—	—	—	—	—	—	2.31	2
poetic	5.80	5	—	—	—	—	6.76	4	5.97	3
touch	5.71	8	—	—	—	—	—	—	—	—
thought	5.67	23	−0.01	5	−0.53	3	1.09	6	2.57	12
images	5.54	11	—	—	2.35	4	0.68	2	4.48	8
exited	5.40	3	—	—	—	—	—	—	—	—
painful	5.35	7	—	—	—	—	—	—	—	—
states	5.20	6	—	—	2.07	2	—	—	1.48	2
mere	5.18	15	−0.52	2	−0.22	2	1.88	5	5.86	14
love	5.11	30	9.39	35	−1.41	3	−1.31	3	−0.80	7
same	5.05	32	−1.35	5	−1.23	4	−1.50	3	2.33	18
hatred	4.98	4	—	—	—	—	—	—	—	—
mental	4.92	3	—	—	—	—	—	—	—	—
mind	4.85	25	−1.40	3	0.72	7	1.31	8	−0.59	6
heart	4.83	18	5.89	16	3.12	9	—	—	−0.33	4

of unique words shown above. Coleridge's vocabulary obviously does overlap with that of other writers to a considerable degree, and overlaps with the vocabulary he employs when discussing other major topics. The main purpose here was to examine to what extent Coleridge's vocabulary showed systematic changes across the period covered by the available notebooks.

In examining changes in the collocations of emotion words, it was also clear that a collocation span of five words would not be adequate. Coleridge's prose in the notebooks contains some unusual stylistic features. Frequent long sentences are characteristic of formal prose of this period, and the notebooks have their share of these. But Coleridge often put down his notes in a series of extended phrases punctuated by dashes or a slash: in this, Coleridge developed a highly flexible prose medium able to capture his distinctive, interconnected manner of thinking. As Coleridge said of himself, in criticising his manner of speaking, he is one of those

> who use five hundred more ideas, images, reasons &c than there is any need of to arrive at their object . . . my illustrations swallow up my thesis — I feel too intensely the omnipresence of all in each, platonically speaking — (NB.ii.2372)

This suggested that the links within Coleridge's thought would often be missed by setting a collocation span as narrow as five words. For the second part of the study, therefore, the span was set at 15 words each side of a target word. In practice the span was often less than this, since the program that was written to extract the data was designed not to cross note boundaries, and many of the notes are very short; the span was also truncated if two of the target words occurred in close proximity, as explained below. Thus the mean span within which collocations were counted turned out to be 9 words.

To identify which words change significantly in distribution across the notebooks, the first method to be used was change-point analysis. In preparation for this, the notebook texts were divided into ten portions of approximately equal length, in which each portion contained material from one or more years (ignoring uncertainties over the dating of some of the notes at this stage). A count was made of all words collocating with emotion words within a span of 15 words in each portion or period. McKinnon's *TextMap* program for identifying words of "aberrant frequency," ABFREQ, was then used to compile a list of the 60 most significant, non-function words within this group. In this program significance is, once again, measured by *z*-score, by comparing the frequency of the collocating words with their frequency in a control text. A large sample of the notebook texts (amounting to approximately one third of the text) was used as the control text. The frequency data for these 60 words was then cast into matrix form (words x period) and analysed by CHANGEPT (also a TextMap program).[8]

In change-point analysis the matrix is cut at the point at which the differences above and below the cut are maximised. Chi square tests are used to assess significance. The data points contributing to the change point — in the present case, words — are listed by CHANGEPT in order of significance. The algorithm is an iterative one: having cut the matrix into two portions, the process is repeated and a second point found in one of the remaining portions. The process can be repeated several times. At each cut the words that are significant on each side of the cut are listed in order of significance.

In the present case five iterations seemed the most appropriate. The first cut was made between 1805 and 1806 (thus dividing the collocate data into two periods, 1794—1805 and 1806—1819). The second cut occurred between 1810 and 1811, dividing the second period into two further periods, 1806-1810 and 1811—1819. These cuts and the three further cuts made are shown in Figure 1.

The data in the figure is arranged with the cuts staggered from left to right in order to indicate the scope of the words occurring on either side of each cut. Thus, the words above the brace showing the first cut are more frequent in the whole period, 1794—1805. A further cut may qualify this scope, however. The third cut, for example, shows that my and I are more frequent below the cut: thus, these words appear to be of most significance above the first cut and below the third, i.e., during 1803—1805. The words at each cut are also listed in order of significance, reading left to right above

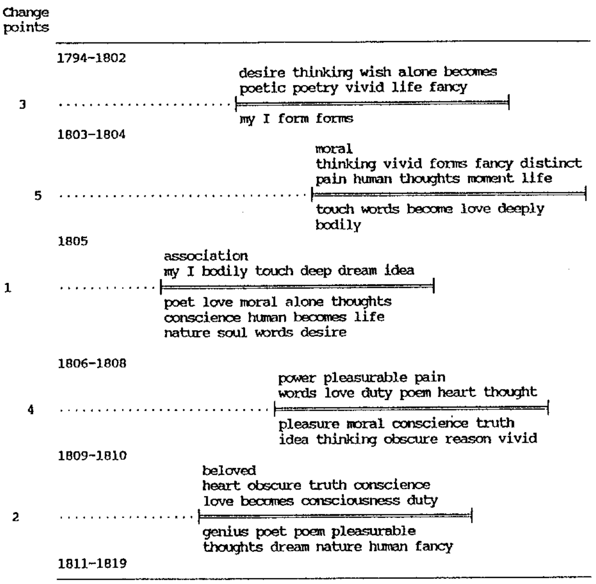
Change
points
1794-1802
desire thinking wish alone becomes
poetic poetry vivid life fancy
3
my I form forms
1803-1804
moral
thinking vivid forms fancy distinct
pain human thoughts moment life
5
touch words become love deeply
bodily
1805
association
my I bodily touch deep dream idea
1
poet love moral alone thoughts
conscience human becomes life
nature soul words desire
1806-1808
power pleasurable pain
words love duty poem heart thought
4
pleasure moral conscience truth
idea thinking obscure reason vivid
1809-1810
beloved
heart obscure truth conscience
love becomes consciousness duty
2
genius poet poem pleasurable
thoughts dream nature human fancy
1811-1819

Figure 1. Change-point analysis of 60 collocates of emotion words in Coleridge's notebooks

or below the brace. Above the first cut, for example, the order of significance is *my, I, bodily, touch, deep, dream, idea, association.*

The analysis shown in Figure 1 provides valuable clues to how Coleridge's concerns evolved across the notebooks, each time he referred to emotion. Looking back in time above the first cut, his concerns in 1805 seem related in particular to his own experience (with self-reference words, my and I, and words reflecting physical experience, *bodily, touch, dream*); but looking further back, several words denoting phenomenal awareness are apparent, *moment, vivid, distinct.* Beyond 1805, however, the words seem to embrace a range of more intellectual or abstract concerns, such as *poet, moral, conscience, human, nature*, and *soul.* This pattern, emerging from the change-point analysis, was then examined in more detail with a second, more fine-grained method of analysis which required the writing of two special programs.

The data contributing to the change-point analysis, it will be recalled, was the frequency with which each word collocated with emotion words at each period. But the significance of the word was its significance as a collocate across the notebooks as a whole, as measured by z-score, which in itself provides no information about changes in the use of the word in different periods. The main limitation of the change-point analysis I have described (and this is how it is intended to be used in conjunction with other *TextMap* programs), is that it takes only the raw frequencies of collocation at each period, ignoring the general pattern of distribution of the word by period. For instance, the word deep may seem to occur as a significant collocate of emotion words in 1805, but the number of times *deep* is used may increase overall in 1805, thus increasing the probability that it will appear in collocation with an emotion word. These considerations suggest that neither the overall significance of a collocation, as measured by z-score, nor the raw frequency of the collocation by period, can in themselves provide an accurate guide to the role of the word when considering changes across time.

Another method was devised, therefore, in an attempt to overcome this difficulty. The words included at this next stage drew upon the collocations found in the z-score lists and on the results of the change-point analysis, but the words were now chosen by estimating their probable significance in the development of Coleridge's thought; and only those words were included which occurred with sufficient frequency across the notebooks as a whole for variance by year to be meaningful.

The notebook files had previously been coded for retrieval in *WordCruncher* using three codes: date, volume and note number, and folio (*WordCruncher* itself was not used for the analysis to be described here). Two new collocation programs were written to take advantage of the date code, which consisted of year, month, and day (if known): this enabled collocation data to be recorded by year, beginning with 1794 and ending with 1819. The dating of the notebooks is not straightforward, however, since some notes cannot be dated with certainty to a given year. The editor of the notebooks, Kathleen Coburn, provides a table of dates for the notes in each volume: this shows certain of the notes as written, according to her best judgement, in one of two, or perhaps three, adjacent years, or in either one of two years separated by an interval (during which it is known that the particular notebook was not used by Coleridge). In such cases, the collocation data was fractioned across the years concerned. For example, if a note was known to have been written in either 1800 or 1812, a collocation occurrence of 0.5 was assigned to both years. The proportion of data falling into this category is small: of the data described below, only 6.6% was of uncertain dating. But the data collected could now be sorted by period in the most accurate form possible.

The first program, COLLOC, read the three volumes of the notebooks and extracted the collocation data. The program was given a selected group of words to find, which were already known to be significant collocates of the target emotion words, as described above. To be included at this stage, however, only those words were chosen that occurred sufficiently frequently to allow comparison of collocates across a range of years; thus words such as *pleasurable, vindictive,* and *pang* were not included. The aim of the study at this stage was to compare the occurrence of these words overall per year with their occurrence in collocation with the target words. Thus the program collected data on both types of occurrence, and recorded it by year. At the same time, with each occurrence of a target word, the actual number of words within the search span was recorded (this was often less than the span of 15 set at the beginning of the program, as noted above).

The words to be found by the program were read into memory already sorted into groups. For example, it was known that *deep* and *deeply* were significant collocates from the *z*-score study. These words were placed in one group, together with other forms of the same word, such as *deeper* and *deepest,* so that occurrences of all forms of the word were placed in one data set. Similarly, *associate, associated, associating, association,* and *associations* were all placed in the group *assoc.* A number of such word groups were examined in this way, until after some trial and error the words in the 24 groups shown below were selected for detailed study.

One other feature of the program should be mentioned, which affected the data being collected. Where target words cluster near to each other, what is to be counted as a collocation? A decision on this is a balance among the alternative possibilities. I decided that a collocate appearing between two closely occurring target words should count only once; but that a collocate occurring to the right of a first and a second target word would count twice. This is a compromise with which I was not entirely happy, but it enabled the density of collocating occurrences to be retained to some degree, without seriously skewing the data for the incidence of collocates.

The occurrence data from COLLOC was saved in a file which was then read by the next program, COLLREAD. The second program also reads in a file containing data on the number of words written in the notebooks per year (data that, again, takes account of uncertainties in dating). Having read the data, the program computed a *z*-score for each word group, taking the text as a whole. This provided a confirmation of the overall significance of the word group in collocating with the target words. This initial report on the word groups studied in detail is shown in Table 2 (note that the *z*-score for *assoc* of 6.82 is for five grouped words, and thus differs from the score of 7.94 shown for the single word *association* in Table 1). The search for collocates was based on the ten emotion target words used in the first study: *feel, feels, feeling, feelings, felt, emotion, emotions, passion, passions,* and *passionate* (these words, it will be recalled, occur a total of 1046 times).

But the main purpose of the COLLREAD program is to provide two sets of percentages: what percentage of the total number of words per year is constituted by the collocate words; and what percentage of all words occurring within the search span are members of each word group. The percentages enable more accurate comparisons of word frequency each year and across years. Percentages also overcome disparities in the number of words written in different years: for example, in 1796 the number of words written was only 5269; in 1804, by contrast, there were 55891. The program, however, reports both raw and percentage data. An example data set is shown in Table 3, based on the word group body (*body, bodily, bodies*), where the number of words per year within the search span and the overall number of words per year is also given.

TABLE 2
Z-scores of selected word groups collocating with emotion words in Coleridge's notebooks

1	mind	+5.50	9	touch	+4.02	17	heart	+4.57
2	idea	+4.02	10	sense	+8.21	18	lang	+2.40
3	distinct	+5.43	11	pain	+4.23	19	life	−0.24
4	assoc	+6.82	12	body	+5.55	20	nature	+4.76
5	vivid	+4.80	13	consc	+4.34	21	poem	+3.37
6	dim	+2.45	14	love	+5.69	22	imagin	+4.49
7	obscure	+7.95	15	duty	+3.15	23	symbol	+0.29
8	deep	+8.82	16	eye	+0.12	24	truth	+3.47

TABLE 3
Frequency and collocation data for the *body* word group in Coleridge's notebooks, 1794—1819, with number of words per year in collocation span and overall number of words per year

	Raw Frequencies		Proportional			
Year	All	Coll	All	Coll	Span words	All words
1794	—	—	—	—	0.0	904
1795	1.0	—	0.100	—	9.5	1005
1796	1.0	—	0.019	—	147.5	5269
1797	—	—	—	—	12.0	2644
1798	1.0	—	0.016	—	28.0	6369
1799	3.0	—	0.017	—	124.7	17973
1800	3.0	—	0.016	—	354.7	19345
1801	5.5	—	0.059	—	395.7	9311
1802	10.5	—	0.052	—	262.0	20092
1803	27.0	5.0	0.066	0.24	2080.0	41061
1804	26.0	3.0	0.047	0.11	2753.0	55891
1805	43.0	9.0	0.106	0.33	2765.0	40492
1806	4.0	1.0	0.029	0.16	630.0	13777
1807	17.0	3.0	0.106	0.34	878.0	16075
1808	14.5	—	0.055	—	1600.3	26523
1809	14.5	—	0.048	—	1118.3	30373
1810	60.5	5.0	0.102	0.20	2540.3	59200
1811	19.0	1.0	0.121	0.08	1233.8	15753
1812	1.0	—	0.017	—	277.0	5843
1813	0.3	—	0.020	—	74.7	1636
1814	1.8	—	0.038	—	117.7	4780
1815	4.3	1.0	0.045	0.57	176.7	9637
1816	1.5	1.0	0.040	3.45	29.0	3789
1817	6.5	—	0.038	—	349.0	17084
1818	23.0	2.0	0.091	0.39	518.5	25199
1819	—	—	—	—	337.0	7040

The percentage data can then be graphed: this helps to reveal potentially significant differences between the two percentages of the collocating word group. The graph for *body* is shown in Figure 2.

The assumption behind this approach is that a word occurring during a given year at approximately the same percentage frequency in collocation as in its general distribution is not a significant collocate of the target words: in such a case, the word forms part of the more general vocabulary employed in the text. This is the case for *body* across much of the notebook period. However, where the word shows a higher percentage for a certain period in collocation than in general, as *body* does for several years, this is a signal that the word might have a distinctive collocational status during that period. The graph suggests that from 1803 through 1807, Coleridge attended particularly to bodily experiences in relation to feelings and emotions — experiences that may have formed a significant element in the development of Coleridge's thought about emotion.

The main limitation of this approach to collocation is if occurrences per year are very small: for example, while there is only one occurrence of a *body* word for 1815, it happens to collocate with a keyword. Being based on only one occurrence, the percentage 0.566 cannot be relied on to establish *body* as a significant collocation of feeling in that year, as the graph seems to indicate (the same problem occurs in more extreme form in 1816). More reliance can be placed on data where three or four years show a similar profile, as with *body* in 1803—07, or where a word with similar function shows a similar profile (as is the case, for instance, when *pain* is compared with *body*). The smaller the number of occurrences, in other words, the less reliance can be placed on the findings (at least, without other supporting evidence).[9]

Examination of the data for a range of word groups allows changes in the primary collocations of the target words to be detected. To facilitate comparison of several word groups, the program also enables changes to be graphed in a set of histograms. Here, it is the difference between the two percentages (where this is positive) that is shown. The graph in Figure 3 represents data for ten of the word groups between 1799 and 1813. These ten word groups best represent the progression of Coleridge's discourse about emotion. Where two or more word groups show similar distributions, the program allows their data to be averaged in one histogram, as has been done here, for example, for the word groups *distinct* and *deep*. The word groups in the figure have been ordered from left to right to indicate some of the changes in collocation across the years. It should be noted that the words may occur as collocates in years that appear blank for that word group: the histogram records only those collocations that are distinctive to a given year.

The substantive result of the technique in the present case, is to bring into focus the phases of Coleridge's development of a view of emotion. This can be broadly summarized by saying that it begins with the phenomenal, moves through a concern for the processes of emotion (as revealed by Coleridge's awareness of changes in consciousness and in the body produced by feeling), then develops into a more systematic and philosophical account of the place of emotion in relation both to an understanding of nature and to Coleridge's views about poetry and poets. These developments can be seen more clearly by reference to a few of the notebook entries concerned — given the large body of material, only a few illustrations and brief comments will be given.[10]

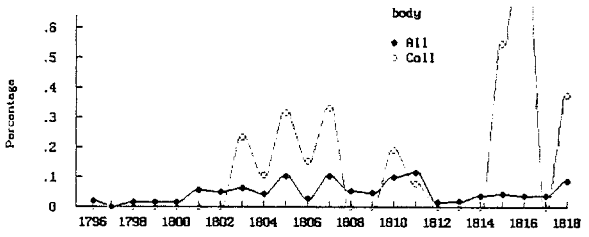

Figure 2. Graph of *body* word group in Coleridge's notebooks, showing percentage frequency of all occurrences by year and occurrences collocating with emotion words

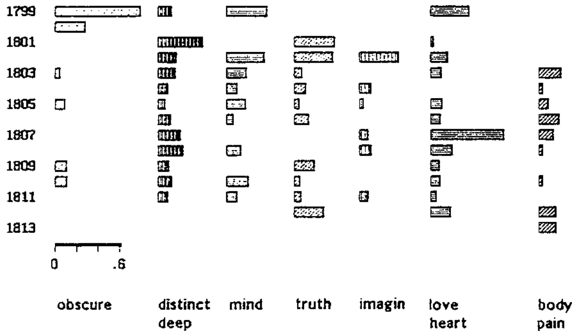

Figure 3. Chart of main collocates of emotion words in Coleridge's notebooks between 1799 and 1813, showing percentage difference between all occurrences and collocating occurrences (where difference > 0)

Phase One: phenomenal

Coleridge noted in 1799 that the best poetry works by suggestion, a phenomenon in which the feelings play a key role (note also the use here of the words *obscure* and *associate*):

> Feelings created by obscure ideas associate themselves with the one *clear* idea. When no criticism is pretended to, & the Mind in its simplicity gives itself up to a Poem as to a work of nature, Poetry gives most pleasure when only generally & not perfectly understood. (NB.i.383; 1799)

This passage also suggests that feelings are brought into play as a result of ideas, a notion that occurs elsewhere: eg., "Strength of Feeling [is] connected with vividness of Idea" (NB.i.1099; 1802); although the opposite phenomenon is also noted: "By deep feeling we make our Ideas dim — & this is what we mean by our Life — ourselves" (NB.i.921; 1801). But in Coleridge's attention to the data of his own consciousness, he increasingly became aware of the role of feelings in mental activity. A key note in this respect is the following, first recorded in 1799 after his first visit to the Hutchinsons (he was to endure a hopeless and painful love for Sara Hutchinson for the next ten years). Coleridge was thinking of the associations of a particular picture that recalled this crucial episode:

> viewed in all moods, unconsciously distinctly, semiconsciously, with vacant, with swimming eyes — a thing of nature thro' the perpetual action of the Feelings! — O God! when I now think how perishable Things, how imperishable Ideas — what a proof of My Immortality — (NB.i.576; 1799)

The feelings, in other words, serve to recall a range of memories that for all intents and purposes seem to be immortal, despite our conventional beliefs about forgetting.

Phase Two: process

The note on memory enables the next phase of Coleridge's ideas about emotion to be seen more clearly, since the note was copied into another notebook four years later with a significant addition:

> O Heaven when I think how perishable Things, how imperishable Thoughts seem to be! — For what is Forgetfulness? Renew the state of affection or bodily Feeling, same or similar — sometimes dimly similar/ and instantly the trains of forgotten Thought rise from their living catacombs! (NB.i.1575; 1803)

Here the process of recall initiated by feeling is made to include "bodily" feeling. By now Coleridge was beginning to take a closer focus on the role of bodily states, and to surmise that changes below the level of awareness are induced by feeling. In this respect, Coleridge was perhaps the first thinker explicitly to postulate a dynamic unconscious. For example:

> Some painful Feeling, bodily or of the mind/ some form or feeling has recalled a past misery to the Feeling & not to the conscious memory — (NB.i.1601; 1803)
>
> — For a Thing at the moment is but a Thing of the moment/it must be taken up into the mind, diffuse itself thro' the whole multitude of Shapes & Thoughts, not one of which it leaves untinged — between wch & it some new Thought is not engendered/this a work of Time/but the Body feels it quicken with me — (NB.i.1597; 1803)

The effects of this understanding can also be traced in Coleridge's account of poetry, which now (in contrast to the note of 1799) gives feelings the primary role:

> Poetry a rationalized dream dealing [?about] to manifold Forms our own Feelings, that never perhaps were attached by us consciously to our own personal Selves . . . — O there are Truths below the Surface in the subject of Sympathy, & how we become that which we understandly [sic] behold & hear, having, how much God perhaps only knows, created part even of the Form. — (NB.ii.2086; 1804)

In this phase the collocation data show that Coleridge is attending in particular to aspects of mental life that indicate formative processes of emotion: states of consciousness, dreams, bodily experience, and specific experiences of pain and misery.

Phase Three: systematizing

The third phase shows the beginning of Coleridge's attempts to systematize what he understood about emotion. This phase clearly continued into the second decade of the 19th century and beyond, and took many forms. One example will suffice to suggest how far Coleridge had moved by

1807. It involves an example of the word *love*, which clearly played a key role in Coleridge's development of a more systematic understanding of emotion. Here the main impetus to understanding was his love for Sara Hutchinson, but the note is remarkable for delineating two realms of consciousness and for its claim that the feelings arbitrate our understanding of their relationship. The note begins with questions about the relation of duty and inclination, and continues:

> The necessary tendency of true Love to generate a feeling of Duty by increasing the sense of reality, & vice versa, a feeling of Duty to generate true Love. All our Thoughts all that we abstract from our consciousness & so form the Phaenomenon Self is a Shadow, its whole Substance is the dim yet powerful sense that it is but a Shadow, & ought to belong to a Substance/but this Substance can have no marks, no discriminating Characters, no hic est, ille non est/it is simply Substance — & this deepliest felt during particular phaenomena with a consciousness that the phaenomenon is in us but it not in the phaenomenon, for which alone we yet value the phaenomenon, constitutes the craving of True Love. (NB.ii.3026; 1807)

In Coleridge's thought an important part is played by discussions of love, but ideas about language, nature, and poetry also play a significant role from 1807 through to the publication of the *Biographia Literaria*. As noted earlier, however, the formulations at which Coleridge arrived in the *Biographia* at times leave emotion out of account. A study of the collocations of emotion words in the notebooks helps to show how important Coleridge's thought about emotion was in developing his mature literary and philosophical theory.

Conclusions

The technique that I have described for mapping changes in collocation around a set of target words is based on a familiar premise. A collocate that occurs more frequently than expected within a set span of the target word is likely to be playing an important role in the usage of that word. In this respect the approach supplements existing statistical techniques, such as the study of z-scores described earlier. The advantage of the present technique is that shifts and changes in a range of collocates across a text can be rapidly surveyed, and portions of the text relevant to a particular collocation identified immediately by use of a simple line graph. Used in conjunction with a program for online text retrieval to check the validity of the findings (*WordCruncher* was the main program used during the present study), the technique also recommends itself for use with students of literature, one of whose concerns will be to look for developments or changes in themes across a text or group of texts. The technique would also be of use in studying a set of verbal protocols in machine readable form, where there is a need to identify contrasts in the collocational environment of key terms between texts. Such applications, however, would require the development of a more flexible and general purpose set of programs than those discussed here, which were created specifically to study Coleridge's notebooks.

Notes

[1] I am grateful to Princeton University Press for permission to use the notebooks in this form.

[2] Among recent studies of changes in content across a text or range of texts, Mark Olsen used the z-score, in "The Language of Enlightened Politics: The *Société de 1789* in the French Revolution," *Computers and the Humanities*, 23 (1989), 357—364; Alastair McKinnon's studies of Kierkegaard have used z-scores, but other measures more recently, such as correspondence analysis: see "Mapping the Dimensions of a Literary Corpus," *Literary and Linguistic Computing*, 4 (1989), 73—84; Nancy Ide used time series and Fourier analysis to discover image patterns in a large Blake poem, in "A Statistical Measure of Theme and Structure," *Computers and the Humanities*, 23 (1989), 277—283.

[3] Samuel Taylor Coleridge, *Collected Letters*, 6 vols., ed. E. L. Griggs, Oxford: Oxford University Press, 1956—71. Davy's letter is cited in Vol. II, p. 1103n.

[4] *The Notebooks of Samuel Taylor Coleridge*, 4 vols., ed. Kathleen Coburn, London: Routledge and Kegan Paul, 1957—90. This and subsequent references are to volume and note number, and are placed in the text. Some notes are quoted in slightly amended form to make them more readable.

[5] Samuel Taylor Coleridge, *Biographia Literaria*, ed. James Engell and W. Jackson Bate, London: Routledge & Kegan Paul, 1983. Vol. II, p. 13.

[6] Godelieve L. M. Berry-Rogghe, "The Computation of Collocations and their Relevance in Lexical Studies," in *The Computer and Literary Studies*, ed. A. J. Aitken, R. W. Bailey, and N. Hamilton-Smith, Edinburgh: Edinburgh University Press, 1974, pp. 103—112. The formula is as follows (I have changed the notation slightly from Berry-Rogghe): P, the probability of collocate B occurring where target word A does not, is given by the formula

$$P = C/(N - A)$$

in which C is the number of observed co-occurrences of B with A, N is the number of words in the text as a whole, and A is the number of occurrences of the target word or words. E is the expected number of co-occurrences, assuming equal distribution of the target and co-occurring words, from the formula

$$E = P \times A \times S$$

where A is the number of occurrences of the target words and S is the span size (here 5 × 2 = 10). The z-score is then given by:

$$z = (K - E)/\sqrt{(E \times Q)}$$

where K is the number of co-occurrences of word B with target words A, and Q = 1 − P. The programs used for this and subsequent analyses were written in compiled BASIC for an 80386 MS-DOS computer; calculation of z-scores was assisted by the word frequency files for the notebooks produced by *WordCruncher* (the *.byf files).

[7] I discussed some of the issues relating to Coleridge's thoughts about love and emotion in "The Aesthetics of Love in Coleridge," *British Journal of Aesthetics*, 23 (1983), 18—24; and "The Displacement of Emotions: the Case of 'Frost at Midnight'," *The Wordsworth Circle*, 20 (1989), 97—102.

[8] Alastair McKinnon, *TextMap* (computer software and manual), Montreal: Inter Editions, 1979. The programs are described in Ian Lancashire and Willard McCarty, eds., *Humanities Computing Yearbook 1988*, Oxford: Oxford University Press, 1988, pp. 330—31. A recent discussion of change-point analysis is: M. S. Srivastava and Keith J. Worsley, "Likelihood Ratio Tests for a Change in the Multivariate Normal Mean," *Journal of the American Statistical Association*, 81 (1986), 199—204.

[9] Another simple way of testing the significance in such cases is the Chi-Square. For example, the data for body words for the central period, 1803—1807, with X^2 (Yates correction), is:

Year	Observed	Expected	X^2
1803	5.0	1.37	7.56*
1804	3.0	1.28	1.22
1805	9.0	2.94	11.33*
1806	1.0	0.18	0.58
1807	3.0	0.93	2.82

* $p = < 0.05$

However, X^2 is less informative than the visual representation provided by a graph, such as that in Figure 2, which gives an immediate sense of the overall trends in the data.

[10] A paper concentrating on the substantive results, "Construing Experience: Coleridge on Emotion," in *The Wordsworth Circle*, 22 (1991), 35—39.