CHATTERTON LECTURE

# POETRY AND THE COMPUTER: SOME QUANTITATIVE ASPECTS OF THE STYLE OF SYLVIA PLATH

By C. S. BUTLER

*Read* 25 October 1979

IF Thomas Chatterton were able to be present here with us this evening, we could, I think, predict a few of the thoughts which might cross his mind. Dominant, perhaps, would be a feeling of indignation that a mere linguist, with no pretensions to special literary sensibilities, should dare to pronounce upon the art form which Chatterton made his own, thereby wasting an hour of his learned audience's valuable time. Mixed with this indignation we should no doubt detect a certain mischievous amusement at the British Academy's apparent decision to celebrate the twenty-fifth year of Chatterton lectures by inviting, in 1979, a namesake of Lionel Butler, who delivered the first lecture in this series in 1955. And undoubtedly our marvellous boy would be unable to suppress a third reaction, one of thankfulness that in his day tonight's lecture, on poetry and the computer, would have been impossible. For there is no doubt that the computer, much used nowadays in authorship attribution studies, would have proved a most tiresome embarrassment for the author of the Rowley poems. It would indeed have been fitting, on the present occasion, to discuss the application of computer techniques to the study of the Chatterton canon. My object this evening is, however, somewhat less ambitious: it is to suggest some of the ways in which the computer can help us to examine certain features of poetic style, and to exemplify these claims from a study of the poetry of Sylvia Plath, an artist who, like Chatterton, suffered a despair which led to suicide.

First, let me say something about the nature of computers, and about their usefulness and limitations in linguistic and literary studies. The computer is a dull and stupid beast, basically responding only to variations in electrical pulses, needing to be led every step of the tortuous way towards a solution to any problem. It is capable of a very few basic activities:

essentially, the recognition of specifically encoded messages, the sorting of data, and the performance of the arithmetical operations of addition, subtraction, multiplication, and division. The computer user, then, must be able to formulate his problem in terms which the machine can recognize, and must set out, step by step, the procedure for solving the problem, again in the computer's terms. Once fed with suitable instructions, however, the machine comes into its own, performing its allotted task with almost unimaginable speed and accuracy.

If, then, the computer is a high-speed moron, what are the implications, for the linguistic and literary researcher, of its potential and limitations? The immense rapidity with which operations are carried out within the machine means that large amounts of textual data can be processed with ease, and their qualitative and quantitative properties examined. For the study of the quantitative properties of texts, it is, of course, advisable to work with sizeable amounts of data in order to obtain statistically meaningful results. Here we see one of the important differences between the work of the computational stylistician and that of the literary critic: the former will tend to make precise statements about a fairly large body of textual data, while the methods employed by the latter allow him to comment not only on a genre, the output of a particular author, or a group of texts, but also on even the shortest poem, a single speech from a play, or a paragraph of narrative description. The emphasis of computational stylistics on the examination of larger textual corpora, and the ability to produce table upon table of figures purporting to describe characteristics of these corpora, bring their attendant dangers. The total frequency of the lexeme *love* in Shakespeare's plays, or the proportion of sentences beginning with καί in the Pauline epistles, are of little interest in themselves. It is only when we begin to compare texts chosen in a principled way, and to relate the facts and figures to an interpretation of those texts, or to problems of authorship, chronology, and so on, that the exercise becomes meaningful.

Given that we are concerned with the comparison of texts in respect of their qualitative and especially their quantitative characteristics, and with the interpretation of our results, we may now ask what kinds of linguistic analysis we are able to perform by computational means. From the pattern of holes punched in a computer card, or from an internal representation of text typed in at the keyboard of a visual display unit,

the machine can recognize elements in its 'character set', consisting basically of the letters of the alphabet, numerals, and punctuation marks, plus certain other symbols, including a space. We see, then, that the computer can recognize what are essentially graphemic elements (i.e. elements which are contrastive in the written language) in a stretch of text; indeed, this is all the machine has to go on when carrying out analyses by means of a program of instructions. It is obvious that the easiest linguistic analyses will be those which operate directly on grapheme sequences. It is, for example, a relatively simple matter to program the computer to count the frequencies of letters or punctuation marks, or to recognize and count the occurrences of particular word forms, regarded as unique sequences of graphemic elements.

If we wish to examine other levels of language—phonology, syntax, lexis, or even semantics or discourse organization—we are faced with serious problems. It is, of course, possible for the researcher to perform his own analyses and encode these together with the text when preparing the data for the machine. He could, for instance, mark each text word with a code indicating its syntactic category, and then simply use the computer to count the occurrences of whatever categories are of interest to him. Such procedures take the drudgery out of sorting and counting operations, but are otherwise of little interest. Rather more exciting are studies using text input with little or no pre-editing, which can be prepared by a typist with no linguistic training, and which use the resources of the computer more fully. The success of such studies will depend on how precisely we can formulate the rules linking the graphological level to the other levels, for the language concerned. At the phonological level, for example, we have the problem of formulating phoneme–grapheme correspondences, in order to obtain information on phenomena of literary interest, such as alliteration. At the lexical level we face the difficulty of 'lemmatization'—the bringing together, under one 'lemma' or vocabulary item, of all the forms which can be taken by that item, for instance all the forms of a verb differing in person, number, tense, and so on—and the related problems caused by homography. The problems involved in syntactic analysis are even greater: here, we must program the machine to deduce, from the sequence of letters, punctuation marks, and spaces, likely syntactic categorizations of textual elements. And if syntactic analysis is difficult, how much more daunting the

prospect of semantic investigations, which are, after all, of primary interest to the seeker of literary meanings. The difficulties of automatic analysis, then, are formidable; nevertheless, considerable progress has been made in recent years, and applications to a variety of literary texts are beginning to bear fruit.

I have dwelt at some length on these general matters because I feel it is all too easy either to claim too much for computational stylistics or to dismiss it as having nothing of serious interest to offer. The truth lies, as so often, in the middle way: in its present state of development, the computational analysis of style is indeed somewhat crude, and has most to say about those linguistic areas which might seem to be of least interest to the more traditional student of literature. And yet, as I hope to demonstrate in what follows, even relatively simple analyses of an essentially graphological nature can provide illuminating insights into the stylistic mechanisms at work in the construction and interpretation of literary texts.

Let us then turn at last to the work of the American poet Sylvia Plath, and see what the computer can tell us about its stylistic development. I shall not talk at length about the life and preoccupations of the poet: a considerable body of biographical and critical material is now available.[1] Some brief introductory remarks are, however, necessary.

Sylvia Plath is of particular interest because her major published poetic works, although spanning a creative period of only about seven years up to her death in 1963, nevertheless show very marked stylistic development. Certain themes recur throughout the poetry: the poet's sense of loss at her German father's death, and her struggle to come to terms with the ambivalence in her relationship with him; her identification with the world of nature which, however, strives to reject her; her own role as mother and wife (she married the poet Ted Hughes in 1956), and as a human being reacting against the horrors of the recent war, horrors to which her German family background made her particularly sensitive; her preoccupation

[1] See, for example, Eileen Aird's *Sylvia Plath* (Edinburgh: Oliver & Boyd, 1973); the collection of articles in Charles Newman (ed.), *The Art of Sylvia Plath: a symposium* (London: Faber & Faber, 1970); Judith Kroll's *Chapters in a Mythology: the Poetry of Sylvia Plath* (New York: Harper & Row, 1976); Edward Butcher's *Sylvia Plath: Method and Madness* (New York: Seabury, 1976) and his *Sylvia Plath: The Woman and the Work* (London: Peter Owen, 1979).

with death; and, especially, her role as artist, painfully shaping poems from the at times frustratingly intractable substance of words.

The first volume of poetry, *The Colossus*,[1] appeared in 1960, and was the only collection to appear in the poet's lifetime. The poems, written between 1956 (when Sylvia Plath was 24) and 1959, concentrate on the poet's relationship with the external world. Rarely, however, even in these early poems, do we find pure description: almost always, the artist's external perceptions are related to an inner conflict, although the fusion of these elements first becomes readily apparent only in the last and longest poem in the collection, 'Poem for a Birthday', which can be seen as a bridge between the other works in *The Colossus* and the later poetry. Stylistically, the poems of *The Colossus* are marked by a rather self-conscious attentiveness to matters of technique. Here we have a young poet of considerable potential, who is flexing her muscles, stretching her technical resources to the utmost, imposing severe limitations of form upon her material. The effect on the reader is of a collection which certainly has something of value to say, but is remarkable chiefly for the ingenuity of its craftsmanship. A poem such as 'Sow', with its mixture of archaisms and colloquialisms, its doubly-hyphenated words, and its complexity in the relationship between grammatical patterning and metrical design, gives a particularly strong sense of the almost artificial, studied cleverness of this early work.

*Sow*

God knows how our neighbour managed to breed
His great sow:
Whatever his shrewd secret, he kept it hid

In the same way
He kept the sow—impounded from public stare,
Prize ribbon and pig show.

But one dusk our questions commended us to a tour
Through his lantern-lit
Maze of barns to the lintel of the sunk sty door

To gape at it:
This was no rose-and-larkspurred china suckling
With a penny slot

[1] The editions used for this study were the Faber & Faber editions published in 1972 (*The Colossus*), 1975 (*Crossing the Water*), 1968 (*Ariel*), and 1975(*Winter Trees*) respectively.

For thrifty children, nor dolt pig ripe for heckling,
About to be
Glorified for prime flesh and golden crackling

In a parsley halo;
Nor even one of the common barnyard sows,
Mire-smirched, blowzy,

Maunching thistle and knotweed on her snout-cruise–
Bloat tun of milk
On the move, hedged by a litter of feat-foot ninnies

Shrilling her hulk
To halt for a swig at the pink teats. No. This vast
Brobdingnag bulk

Of a sow lounged belly-bedded on that black compost,
Fat-rutted eyes
Dream filmed. What a vision of ancient hoghood must

Thus wholly engross
The great grandam!—our marvel blazoned a knight,
Helmed, in cuirass,

Unhorsed and shredded in the grove of combat
By a grisly-bristled
Boar, fabulous enough to straddle that sow's heat.

But our farmer whistled,
Then, with a jocular fist thwacked the barrel nape,
And the green-copse-castled

Pig hove, letting legend like dried mud drop,
Slowly, grunt
On grunt, up in the flickering light to shape

A monument
Prodigious in gluttonies as that hog whose want
Made lean Lent

Of kitchen slops and, stomaching no constraint,
Proceeded to swill
The seven troughed seas and every earthquaking continent.

If we now turn to the poems written towards the end of Sylvia Plath's life, and collected, after her death, by her husband in the volumes *Ariel* and *Winter Trees*, we witness a withdrawal into the innermost depths of a tortured personality. The sense of desolation conveyed so starkly by the poet is made even more frightening by being set against the familiar background of domestic routine and other workaday events. Stylistically, the

difference from the early poems of *The Colossus* is remarkable. Gone is the mannered turn of phrase, the studied avoidance of the ordinary; in their place we find an intense concentration of imagery, combined with a greater simplicity of linguistic form. The graphology, the phonology, and syntax of the poems no longer attract so much of our attention; on the other hand, the complexity and power of the semantics (and who knows what higher levels?) is spell-binding. The tone has become much more conversational, more colloquial. Sylvia Plath herself said[1] that any lucidity her later poems might possess came from the fact that she said them to herself aloud. The following short poem from *Ariel* is not untypical of this later style.

*Contusion*

Colour floods to the spot, dull purple.
The rest of the body is all washed out,
The colour of pearl.

In a pit of rock
The sea sucks obsessively,
One hollow the whole sea's pivot.

The size of a fly,
The doom mark
Crawls down the wall.

The heart shuts,
The sea slides back,
The mirrors are sheeted.

The fourth volume of poetry which we shall discuss here, *Crossing the Water*, has been seen by critics as transitional between *The Colossus* and the later work. The poems, which were prepared by Ted Hughes and first published in collected form in 1971, belong chiefly to the period from late 1959 to early 1962 and reflect, among other things, the poet's internal reactions to the major events in her life at this time—the move from America to England and the birth of her two children. As Eileen Aird has said:[2]

> As an experimental, transitional volume *Crossing the Water* is very valuable as a bridge between the early composure of *The Colossus* and the later originality and daring of *Ariel*.

In our all-too-brief discussion of the development of Sylvia

[1] Peter Orr (ed.), *The Poet Speaks* (London: Routledge & Kegan Paul, 1966), p. 170.

[2] Op. cit., p. 50.

Plath's poetic art we have already formulated, in a very subjective way, a rather general hypothesis about the evolution of her style: namely, that while the earlier poems are somewhat contrived and artificial in their language, with considerable complexity and subtlety of linguistic form, the later poems are much more conversational, the complexity being more semantic than formal.[1] I shall now try to give sharper definition to these somewhat vague suggestions by putting forward a number of more specific hypotheses, which are derivable from our general view, and which are capable of being tested by a computational analysis of the texts.

For any text, or set of texts, we may arrive at various measures of vocabulary richness by counting the total number of word-tokens, the number of different words ('types'), and the distribution of word-frequencies. Such information is easily obtained by means of one of the 'package' programs now available. These have the advantage of being designed for use by researchers who are not specialists in the computer field. No knowledge of special programming languages is required, the instructions to the program being phrased in terms sufficiently similar to English for the naïve user not to feel too uncomfortable. The package used for part of the present investigation was COCOA (word *CO*unt and *CO*ncordance on *A*tlas), originally designed at the Atlas Computer Laboratory, which can produce word-frequency lists arranged in various ways, and concordances or indexes of word-occurrences.

One measure of vocabulary richness is the 'type–token ratio' (TTR). This is the ratio of the number of different words (i.e. the 'vocabulary' of the text) to the total number of running words. The TTR is most valuable when calculated on the basis of lexemes (words in the dictionary sense) as types. This, however, involves the problem of reuniting under one head all the forms which a given lexeme can take, and we shall use word-forms rather than lexemes here, as has been done in many other studies. Unfortunately, the TTR is not independent of text length, and so cannot validly be used to compare the vocabulary richness of texts of widely differing lengths. Various attempts have been made to develop a measure of vocabulary richness which is independent of text length within wide

[1] By this I do not, of course, mean to imply that conversational language shows no formal complexity. The complexity is, however, of a rather different kind from that found in stylized written language of the type found in *The Colossus*.

limits. One formula which, it is claimed, satisfies these conditions is the following, developed recently by Honoré:[1]

$$R = \frac{100 \log N}{1-(V_1/V)}$$

where: $R$ = vocabulary richness measure

$N$ = total number of word tokens in text

$V$ = number of different word types in text

$V_1$ = number of words used once only

A slightly different approach to vocabulary richness is to estimate the expected reduction in vocabulary (word-types) for a given reduction in total words (tokens), keeping the vocabulary richness constant. This allows us to predict the number of types in a shorter text from the number present in a longer text, on the assumption of equal richness. We can then compare the predicted vocabulary with the actual vocabulary. The rationale behind this procedure is quite simple. Let the total number of word-tokens in a text be $N$, and the number of different word-types $V$, and let the number of word-types occurring once, twice, three times, and so on, be $V_1$, $V_2$, $V_3$, etc. Now let us imagine that we shorten the text so that it now has only $M$ tokens, where $M<N$. The question is how many different words we may expect in our new shortened text. The text has been reduced by a proportion $(N-M)/N$. Considering a word-type appearing only once (a 'hapax legomenon'), it is clear that the probability of such a word having its sole occurrence in the part of the text which has been removed is $(N-M)/N$. Since the number of hapaxes in the original text was $V_1$, we may expect a reduction of $V_1(N-M)/N$ from this source. Now consider the words appearing twice in the original text. The probability that *both* occurrences are in the deleted part of the text is $([N-M]/N)^2$, and the expected reduction is $V_2([N-M]/N)^2$. We can go on in this way and sum the expected reductions. In practice, if the text is shortened by a relatively small proportion, the reductions become very small above a power of two, and it is not normally necessary to proceed further.

What, you may well ask, has all this mathematical juggling to do with the poetic genius of Sylvia Plath? We suggested earlier that the early poems collected in *The Colossus* paid

[1] Tony Honoré, 'Some simple measures of richness of vocabulary', *ALLC Bulletin*, 7 (2), 1979, pp. 172–7.

scrupulous attention to matters of technique, avoiding the banal at all costs. We might, then, expect that the vocabulary richness of these poems will be significantly higher than that of the later poems in *Ariel* and *Winter Trees*, with *Crossing the Water* as a transitional volume with intermediate values of richness. Table 1 shows the relevant primary data, also the value of TTR and Honoré's $R$, for each of the four volumes of poems, a word being defined here as any sequence of letters, hyphens, and apostrophes bounded by spaces.

TABLE 1

*Word frequency data*

| | *Colossus* | *Crossing the Water* | *Ariel* | *Winter Trees* |
|---|---|---|---|---|
| Total number of word tokens ($N$) | 8827 | 7949 | 9340 | 6100 |
| Number of different word types ($V$) | 3221 | 2560 | 2571 | 1898 |
| Number of words occurring once ($V_1$) | 2360 | 1807 | 1741 | 1254 |
| Number of words occurring twice ($V_2$) | 420 | 336 | 344 | 272 |
| Proportion of word types occurring only once ($V_1/V$) | 0.733 | 0.706 | 0.677 | 0.661 |
| Type-token ratio ($V/N$) | 0.365 | 0.322 | 0.275 | 0.287 |
| Honoré's $R = \frac{100 \log N}{1 - \frac{V_1}{V}}$ | 1478 | 1327 | 1229 | 1117 |

The TTR is somewhat unreliable here, in view of the considerable differences in text lengths; it is reassuring, however, that it varies in the predicted direction. Honoré's $R$, which as we have seen is independent of text lengths, shows a steady drop from *The Colossus*, through *Crossing the Water*, to *Ariel* with *Winter Trees* having the lowest value of all. These results bear out our hypothesis in a very convincing way.

Using the data in Table 1, and the calculation for which we earlier provided a rationale, we can estimate the expected decrease in vocabulary from one of the longer texts to one of the shorter ones, assuming equal vocabulary richness. The results for each possible pairing of texts are shown in Table 2.

It is clear that the actual vocabulary in *The Colossus* is considerably greater than would be predicted on the basis of the vocabulary distribution in *Ariel*, while that of *Winter Trees* is more restricted than a forecast from the other three texts might suggest. *Crossing the Water* does indeed seem to be intermediate between the two extremes, its vocabulary being richer than that of *Ariel*, but poorer than that of *The Colossus*.

TABLE 2

*Prediction of vocabulary in shorter texts from that in longer texts, given constant vocabulary richness*

| | | Vocabulary expected in | | | |
|---|---|---|---|---|---|
| | | *Colossus* | *Crossing the Water* | *Ariel* | *Winter Trees* |
| As predicted from | *Colossus* | — | 2982 | — | 2602 |
| | *Crossing the Water* | — | — | — | 2246 |
| | *Ariel* | 2474 | 2314 | — | 2033 |
| | *Winter Trees* | — | — | — | — |
| | Actual vocabulary | 3221 | 2560 | 2571 | 1898 |

A second area of investigation facilitated by computational techniques is that of punctuation. We need to be slightly wary here: as I mentioned earlier, only one of the collections was published under Sylvia Plath's own supervision, and editors of manuscripts are notorious for taking liberties with the author's punctuation. Perhaps we may be somewhat reassured, however, by the fact that at least some of the middle and late poems had appeared singly in reviews and magazines before the poet's death (although this obviously does not preclude the possibility of differences between magazine-form and volume-form), and by the knowledge that the posthumous volumes were put together by Sylvia Plath's husband, who had obviously been closely associated with the writing of the poems. I remarked earlier that in the poems of *The Colossus*, Sylvia Plath appeared to use rather complex grammatical patterns. The impression given is of the conscious avoidance of congruence between lines and syntactic divisions, resulting in a low proportion of end-stopped lines. In the later poems of *Ariel* and *Winter Trees*,

however, we often find simple sentences occupying a single line, or more than one whole line. These impressions are borne out by the data in Table 3, obtained using a computer program specially written in the language SNOBOL 4, whose string-handling and pattern-matching facilities make it particularly suitable for linguistic and literary computing.

TABLE 3

*End-stopping*

| | *Colossus* | *Crossing the Water* | *Ariel* | *Winter Trees* | Total |
|---|---|---|---|---|---|
| Number of end-stopped lines | 764 (982.54) | 745 (694.00) | 1098 (1023.09) | 740 (647.37) | 3347 |
| Number of run-on lines | 690 (471.46) | 282 (333.00) | 416 (490.91) | 218 (310.63) | 1606 |
| | 1454 | 1027 | 1514 | 958 | 4953 |
| $\chi^2$ (d.f. = 3) = 219.26 | | | $p \ll 0.001$ | | |

Knowing the proportion of end-stopped lines (and therefore also that of run-on lines) in the whole of the corpus, and the length of each set of poems, we can calculate the expected number of end-stopped and run-on lines for each of the four volumes, assuming no differences between the texts in this respect. The expected values are given in brackets in Table 3. The so-called 'chi-square' ($\chi^2$) test allows us to assess the significance of the differences between expected and observed frequencies. The value of $\chi^2$ is given by:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where: $\sum$ = sum over all pairs of expected and observed frequencies

$O$ = observed frequency

$E$ = expected frequency

Statistical tables are available, giving the probability of exceeding a given value of $\chi^2$ by chance, for a particular number of 'degrees of freedom', defined as $(R-1)(C-1)$, where $R$ is the number of rows and $C$ the number of columns in the appropriate 'contingency table' such as is represented by Table 3. In the present case, there are two rows and four columns, so

that the number of degrees of freedom is $(4-1)\times(2-1)$, or 3. The value of $\chi^2$ here is so high that the probability of obtaining it by chance is very small indeed (much less than 1 in 1000). The main contributions to this high value are those outlined in boxes in Table 3: the large excess of run-on lines in *The Colossus* (with a corresponding deficit of end-stopped lines), and the large deficit of run-on lines in *Winter Trees* (with a corresponding excess of end-stopped lines). *Ariel* also has a considerable deficit of run-on lines, as predicted. It is interesting that in respect of this parameter, the 'transitional' text *Crossing the Water* resembles *Ariel* and *Winter Trees* rather than *The Colossus*, although its deviations from the expected values are not large.

Table 4 shows the distribution of the numbers of punctuation marks over the four sets of poems. The expected values are again given in brackets.

TABLE 4

*Total punctuation*

| | *Colossus* | *Crossing the Water* | *Ariel* | *Winter Trees* | Total |
|---|---|---|---|---|---|
| Total number of punctuation marks | 1431 (1576.54) | 1304 (1419.72) | 1783 (1668.16) | 1327 (1180.57) | 5845 |
| $\chi^2$ (d.f. = 3) = 48.94 | | | $p < 0.001$ | | |

Again the value of $\chi^2$ is highly significant, the main contributions being made by the large deficit in *The Colossus* and the large excess in *Winter Trees*. Interestingly, in respect of total punctuation *Crossing the Water* resembles *The Colossus* rather than the later works.

It is also instructive to consider the distribution of individual punctuation marks in the texts. Before we look at the data provided by our punctuation-counting program, let us again attempt to formulate some hypotheses based on our general characterization of the texts. If *The Colossus* has more complex grammetric patterns, while *Ariel* and *Winter Trees* often have simple clauses or whole sentences occupying a single line, we might expect the full stop and comma (the most likely separators of sentences and clauses) to have a higher relative frequency in the later poems. The colon and semi-colon, on the other hand, are suited to formal writing, in which a complex sentence may need to be split up into parts with only minimal interdependence; we might therefore expect these punctuation marks

to be more frequent in *The Colossus* than in *Ariel* and *Winter Trees*. Question marks (which may invite the reader's participation in a pseudo-conversational manner), exclamation marks, dashes, and quotation marks may be taken as indicators of a greater informality of language, appropriate to a conversational style. The data for our four sets of poems are given in Table 5, with $\chi^2$ values for the individual punctuation marks.

TABLE 5

*Individual punctuation marks*

| Punctuation mark | *Colossus* | *Crossing the Water* | *Ariel* | *Winter Trees* | Total | $\chi^2$ | $p$ |
|---|---|---|---|---|---|---|---|
| full stop | 536 (610.93) | 545 (550.16) | 660 (646.43) | 524 (457.48) | 2265 | 19.22 | < 0.001 |
| comma | 749 (761.43) | 581 (685.69) | 885 (805.68) | 608 (570.19) | 2823 | 26.50 | < 0.001 |
| colon | 65 (39.11) | 41 (35.22) | 24 (41.38) | 15 (29.29) | 145 | 32.36 | < 0.001 |
| semi-colon | 25 (17.80) | 19 (16.03) | 16 (18.84) | 6 (13.33) | 66 | 7.92 | < 0.05 |
| exclamation mark | 17 (36.14) | 34 (32.55) | 29 (38.24) | 54 (27.07) | 134 | 39.22 | < 0.001 |
| question mark | 5 (40.73) | 16 (36.68) | 77 (43.09) | 53 (30.50) | 151 | 86.29 | ≪ 0.001 |
| bracket | 2 | 6 | 2 | 0 | 10 | figures too low to apply test | |
| dots (indicating unfinished sentence) | 0 | 2 | 1 | 2 | 5 | " " " | |
| dash | 30 (56.64) | 52 (51.01) | 77 (59.93) | 51 (42.42) | 210 | 19.15 | < 0.001 |
| quotation marks | 2 (9.71) | 8 (8.74) | 12 (10.27) | 14 (7.27) | 36 | 12.71 | < 0.01 |

The data in Table 5 support our hypotheses very strongly. The differences between *The Colossus*, on the one hand, and *Ariel* and *Winter Trees*, on the other, are always in the predicted direction, and in all cases but one, the main contributions to $\chi^2$ are made by *The Colossus* and one or both of the late volumes. The exception is the comma, where the main opposition seems to be between *Crossing the Water*, with a large deficit, and *Ariel*,

with a large excess. In all other cases, except that of the question mark (where large differences from the expected values are shown by all texts), *Crossing the Water* shows frequencies closer to those expected on the basis of the corpus as a whole. Thus, *Crossing the Water* again emerges as a transitional collection, as predicted.

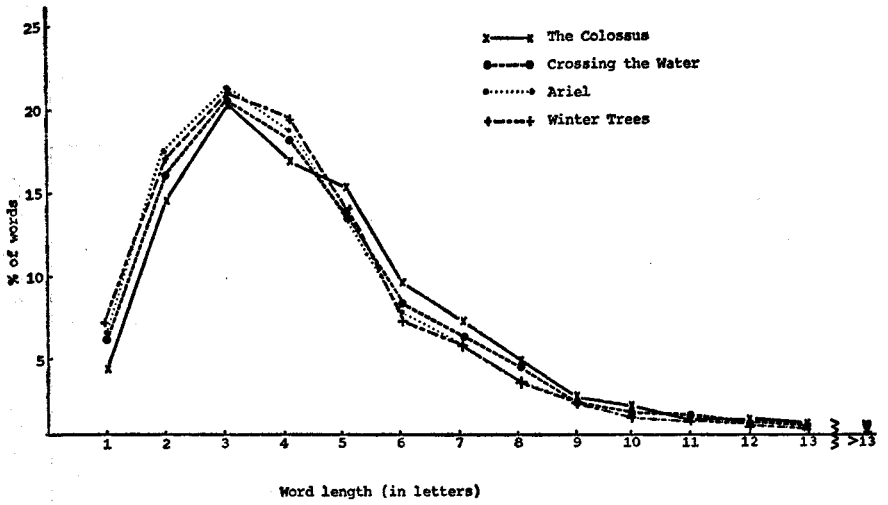


FIG. 1

A further parameter which we might expect to correlate with the opposition between a more formal, elegant, 'clever' style and a more informal and relaxed style is word-length, which, we predict, will on the whole be smaller in the later works than in *The Colossus*. A program was written in SNOBOL 4 to produce word-length profiles for each of the collections of poems. Alphabetic characters, apostrophes, and hyphens counted as 'letters' for the purpose of word-length computation. Figure 1 shows graphically the word-length distribution for each set of poems. The distributions are all of similar shape, and their strongly skewed nature precludes the use of the statistical test known as the '*t*-test' for comparing the mean word-lengths for each set of texts. The modal value (the word-length showing the highest frequency) is 3 letters in each case. It is noticeable that the proportion of words with 1–4 letters is higher in *Ariel* and *Winter Trees* than in *Crossing the Water*, which in turn has a higher proportion than *The Colossus*; while for words of 5 letters or more *The Colossus* has the highest proportion, *Crossing the Water* again being in an intermediate

position for words of 6 and 7 letters. In Table 6 the word lengths are grouped in classes with an interval of 3 letters, and differences in distribution are tested by means of the $\chi^2$ procedure.

Table 6 shows that there are highly significant differences in word-length distribution across the texts. The main contributions to $\chi^2$ are made by the values for *The Colossus* and

TABLE 6

*Word length*

| Word length (letters) | *Colossus* | *Crossing the Water* | *Ariel* | *Winter Trees* | Total |
|---|---|---|---|---|---|
| 1–3 | 3473 (3810.67) | 3417 (3431.63) | 4238 (4032.13) | 3000 (2853.57) | 14128 |
| 4–6 | 3743 (3617.54) | 3221 (3257.71) | 3774 (3827.78) | 2674 (2708.96) | 13412 |
| 7–9 | 1272 (1117.47) | 1042 (1006.32) | 1076 (1182.41) | 753 (836.80) | 4143 |
| 10–12 | 285 (243.83) | 234 (219.58) | 220 (258.00) | 165 (182.59) | 904 |
| > 12 | 54 (37.49) | 35 (33.76) | 32 (39.67) | 18 (28.07) | 139 |
| Total | 8827 | 7949 | 9340 | 6610 | 32726 |
| $\chi^2$ (d.f. = 12) = 122.18 | | | $p \ll 0.001$ | | |

some of those for *Ariel* and *Winter Trees*. *The Colossus* has a much smaller proportion of words in the 1–3-letter range, but a higher proportion of all classes of longer words, than would be expected from an even distribution over all four sets of poems. *Ariel* and *Winter Trees*, on the other hand, have an excess of words in the 1–3-letter range, but fewer words in the mid to high ranges, especially between 7 and 12 letters. Again, *Crossing the Water* has values which approximate more closely to those expected on the basis of the corpus as a whole. The results thus support our hypothesis very strongly indeed. This is all the more remarkable since in all four sets of texts (and indeed in any text in the English language) there is a high proportion of certain short 'grammatical' words, which might have been expected to mask any stylistic effects.

Encouraged by the results of word-length investigations, let us now consider sentence-length. Clearly, since our subjective impressions have led us to claim that *The Colossus* has more

involved linguistic patterning than the later works, we should expect this to be reflected in a higher sentence-length. A SNOBOL 4 program was written to produce sentence-length profiles, a sentence being defined as a stretch of language between full stops, question marks, exclamation marks, or dots indicating an unfinished sentence. Figure 2 shows graphically

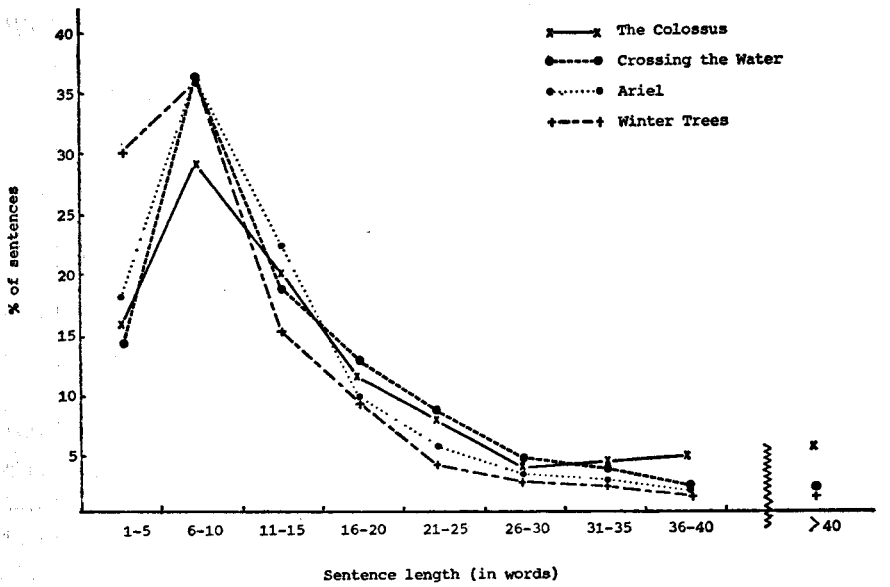


FIG. 2

the sentence-length distribution for each set of texts, the lengths being grouped into classes with a 5-word interval. As with word-length, the distribution is skewed; the modal value in each case lies in the 6–10 word range. The data are presented in tabular form in Table 7, which also shows the values expected if sentence-length were distributed evenly over the four sets of texts.

The $\chi^2$ value is again highly significant, over half of it being contributed by the differences shown in boxes in Table 7. As predicted, *The Colossus* has an excess of longer sentences (especially in the range above 35 words), while *Winter Trees* has a large excess of very short (1–5-word) sentences. *Crossing the Water* is like *The Colossus* with respect to its very short sentences, but behaves in an intermediate way at higher sentence-lengths, having high relative frequencies in the middle range (16–30 words). Contrary to our predictions, however, *Ariel* also behaves in an intermediate fashion—indeed, in respect of its

shortest sentences it resembles *The Colossus* more than *Winter Trees*.

One linguistic feature which strikes the reader of the early poems collected in *The Colossus* is the poet's predilection for the

TABLE 7

*Sentence length*

| Sentence length (words) | *Colossus* | *Crossing the Water* | *Ariel* | *Winter Trees* | Total |
|---|---|---|---|---|---|
| 1–5 | 90 (110.73) | 85 (118.47) | 141 (152.20) | 191 (125.61) | 507 |
| 6–10 | 159 (192.84) | 218 (206.32) | 277 (265.07) | 229 (218.76) | 883 |
| 11–15 | 113 (107.01) | 113 (114.49) | 170 (147.10) | 94 (121.40) | 490 |
| 16–20 | 66 (58.09) | 72 (62.15) | 71 (79.85) | 57 (65.90) | 266 |
| 21–25 | 42 (34.94) | 50 (37.39) | 44 (48.03) | 24 (39.64) | 160 |
| 26–30 | 19 (17.91) | 26 (19.16) | 23 (24.62) | 14 (20.32) | 82 |
| 31–35 | 19 (13.54) | 15 (14.49) | 18 (18.61) | 10 (15.36) | 62 |
| 36–40 | 21 (10.70) | 9 (11.45) | 11 (14.71) | 8 (12.14) | 49 |
| > 40 | 29 (12.23) | 9 (13.08) | 12 (16.81) | 6 (13.87) | 56 |
| Total | 558 | 597 | 767 | 633 | 2555 |
| $\chi^2$ (d.f. = 24) = 134.52 | | $p \ll 0.001$ | | | |

use of hyphenated words. Certain kinds of hyphenated compound, such as those formed from a noun plus an adjective (e.g. *marble-heavy*, *rust-red*, *rabbit-eared*) are particularly suitable as vehicles for compressed similes, and we might expect them to be more frequent in the highly stylized poems of *The Colossus* than in the later more conversational style. Table 8 gives the relevant data, obtained by means of the COCOA package, which can be instructed to search for hyphenated words. The 'expected' values are calculated on the basis of the total number of hyphenated words in the corpus, and the lengths of the four sets of poems.

As predicted, *The Colossus* has a high proportion of hyphenated words, and *Ariel* and *Winter Trees* have a low proportion, while *Crossing the Water* shows approximately the proportion present in the corpus as a whole.

If we now look at the syntactic types of hyphenated compound present in the texts, further interesting differences emerge. There are four main categories of compound, plus a number of minor types. The frequent combinations are: noun

TABLE 8

*Hyphenated words*

| | *Colossus* | *Crossing the Water* | *Ariel* | *Winter Trees* | Total |
|---|---|---|---|---|---|
| Number of hyphenated words | 194 (106.81) | 100 (96.20) | 69 (113.01) | 33 (79.98) | 396 |
| $\chi^2$ (d.f. = 3) = 116.07 | | | $p \ll 0.001$ | | |

plus noun (e.g. *bird-feet, flower-head, bee-seller*); noun plus word ending in *-ed*, which may be a past participle or a denominal adjective (e.g. *gull-coloured, moon-skulled, heat-cracked*); noun plus some other part of speech, often but not always an adjective (e.g. *fever-dry, peanut-crunching*); adjective plus word ending in *-ed*, which may again be a past participle or denominal adjective (e.g. *green-shaded, absent-minded, bent-backed*). Table 9 gives the distribution of these categories across the texts.

It should first be admitted that the $\chi^2$ test is perhaps not quite as reliable here as elsewhere, since strictly speaking it is not advisable to use the test if expected values drop below 5. Nevertheless, there are indications of interesting differences. *The Colossus* appears to have a high proportion of compounds in the 'noun + *-ed*' and 'noun + other' categories, while *Ariel* has a low proportion of these types. It is not entirely clear to me how these findings should be interpreted, although looking through the lists of these types of compound it did seem to me that many of them embody rather striking images, while some of the other types (e.g. noun + noun) lend themselves more readily to somewhat banal, everyday descriptive expressions (e.g. *nose-end, rose-stem, sea-salt, fish-bait*).

Just as hyphenated compounds can tell us something about the style of these poems, so can words containing apostrophes, which can again be searched for using the COCOA program.

TABLE 9

*Syntactic types of hyphenated words*

| Syntactic type | *Colossus* | *Crossing the Water* | *Ariel* | *Winter Trees* | Total |
|---|---|---|---|---|---|
| noun + noun | 80 (79.36) | 37 (40.91) | 30 (28.23) | 15 (13.50) | 162 |
| noun + *-ed* word | 44 (30.37) | 11 (15.66) | 3 (10.80) | 4 (5.17) | 62 |
| noun + other | 11 (18.62) | 11 (9.60) | 11 (6.62) | 5 (3.17) | 38 |
| adj. + *-ed* word | 26 (21.56) | 10 (11.11) | 6 (7.67) | 2 (3.67) | 44 |
| others | 33 (44.09) | 31 (22.73) | 19 (15.68) | 7 (7.50) | 90 |
| Total | 194 | 100 | 69 | 33 | 396 |
| $\chi^2$ (d.f. = 12) = 30.02 $p < 0.01$ | | | | | |

TABLE 10

*Contracted forms*

| | *Colossus* | *Crossing the Water* | *Ariel* | *Winter Trees* | Total |
|---|---|---|---|---|---|
| Number of contracted forms | 53 (53.14) | 87 (47.85) | 31 (56.22) | 26 (39.79) | 197 |
| $\chi^2$ (d.f. = 3) = 48.12 $p < 0.001$ | | | | | |

Apostrophes are used for two basic purposes in English: to indicate possession and to form contractions. It is the latter use which is of interest from the point of view of Sylvia Plath's style, for we might expect that the more conversational poems of *Ariel* and *Winter Trees* would contain a higher proportion of contracted forms than the earlier poems of *The Colossus*. The relevant data are given in Table 10.

Table 10 presents an extremely interesting picture, completely at variance with our predictions. *The Colossus* has just the number of contracted forms expected from an even distribution across the texts; *Crossing the Water* has a very high proportion of such forms; *Ariel* and *Winter Trees* both have low proportions. We may perhaps explain the high value in *Crossing the Water* as indicating an effort, in this transitional

experimental period, to move away from the formality of much of *The Colossus*, by the deliberate use of a feature characteristic of the spoken and informal written languages. Nevertheless, the low proportion of contracted forms in the late poems is surprising, in view of the strong convergence of other factors pointing towards a more conversational style.

TABLE 11

*Syntactic types of contracted forms*

| Syntactic type | *Colossus* | *Crossing the Water* | *Ariel* | *Winter Trees* | Total |
|---|---|---|---|---|---|
| aux. + neg | 15 (17.76) | 35 (29.15) | 7 (10.39) | 9 (8.71) | 66 |
| pron. + aux./*be*/*have* | 18 (25.02) | 45 (41.07) | 18 (14.63) | 12 (12.27) | 93 |
| noun + '*s* | 16 (6.46) | 3 (10.60) | 3 (3.78) | 2 (3.17) | 24 |
| adverb + '*s* | 4 (3.77) | 4 (6.18) | 3 (2.20) | 3 (1.85) | 14 |
| Total | 53 | 87 | 31 | 26 | 197 |
| Certain values too low for $\chi^2$ test to be reliable | | | | | |

In view of this rather unexpected result, let us examine the distribution of syntactic types of contracted forms. There are four major types: auxiliary verb plus negative (e.g. *can't*, *doesn't*); pronoun plus auxiliary verb or part of main verb *be* or *have* (e.g. *she'll*, *I'm*, *it's*, *that's*); noun plus contracted form of *is* (e.g. *bastard's*, *gossip's*); adverb plus contracted form of *is* (e.g. *there's*, *here's*). Table 11 shows the distribution of these types across our texts.

Unfortunately, there are too many low expected values for the $\chi^2$ test to be reliable. The data do, however, indicate a high relative proportion of the 'noun + '*s*' type in *The Colossus*, and a low value for this type in *Crossing the Water*. Thus, although *Crossing the Water* has the highest over-all proportion of contracted forms, it is poor in just that type which is present in *The Colossus* with higher frequency than expected. There is thus an indication (which cannot, however, be statistically firmed up) of a difference, although it is not at all obvious why it should reside in the 'noun + '*s*' type of contraction rather than any other.

x

Let me now try to bring together the various strands of our discussion. We began with the general hypothesis that the poems of *The Colossus* showed a somewhat stilted, formally complex style, while the late poems were more informal and colloquial in tone. We then sharpened this rather vague suggestion by formulating a number of related but more specific sub-hypotheses, each of which could be tested. The computer was not, of course, essential to this testing. With the aid of package programs and a small set of programs specially written in SNOBOL 4, however, it was possible to save the considerable labour and mind-anaesthetizing boredom of a manual analysis, and to achieve a high degree of accuracy in the results. In almost all cases (the notable exception being the distribution of contracted forms) the results of our tests provided strong support for our hypotheses.

The studies I have described here are, of course, merely the tip of the iceberg, since they make direct use of the information obtained from the sequence of graphemes encoded in the text. In conclusion, I should like to mention some avenues which will be explored in further work on Sylvia Plath's poetry. I have said nothing so far about thematic analysis, an area more closely allied to the traditional concerns of the literary critic. I have begun to examine certain semantic areas, chiefly those of colour and family relationships, using word-lists and concordances generated by means of COCOA and similar packages. Preliminary results on the colour field suggest that while the distribution of most of the frequent colour-terms is similar in the four volumes of poems, their collocational behaviour (that is, the kinds of lexeme with which they tend to associate) may vary with the stages in the poet's development. A further area for future investigation is the syntax of the texts. A package developed by Ross and Rasche in the USA (EYEBALL), and modified as OXEYE by Dr Burnard, of the Oxford University Computing Service, for use on the ICL 1900 series of computers, may prove useful here, although preliminary runs suggest that the unusual syntactic patterns often present in poetry may result in a low success rate in automatic syntactic parsing by means of the program. If the British Academy is unable to find another Butler for the 2004 Chatterton Lecture, perhaps I shall be able to let you know.