POPS AND FLOPS: Some Properties of Famous English Poems

Richard S. Forsyth

University of Luton
U.K.

[
Contact:

forsyth_rich@yahoo.co.uk

Cite as:

Forsyth, R.S. (2000). Pops & Flops: some properties of famous English poems. *Empirical Studies of the Arts*, 18(1). 49-67.
]

Abstract

This paper describes a preliminary study of linguistic attributes that differentiate popular from obscure poems in English. Following in the footsteps of Simonton (1989), Martindale (1990) and others, frequency of appearance in anthologies was used as an index of poetic popularity. Twenty general anthologies published between 1966 and 1997 were selected and all poems appearing in more than five of them were taken as a reference sample. This gave 85 poems by 54 different authors. (The two most popular were Matthew Arnold's *Dover Beach* with 16 occurrences and *Kubla Khan* by Samuel T. Coleridge with 15.)

As a control group, 54 other poets were selected by finding a less eminent poet of the same sex born within 10 years of each poet in the reference sample. The same number of poems were chosen (as near as possible randomly) from each obscure poet as from the matching popular poet. This gave 85 obscure poems, also by 54 different authors. As a check on this dichotomy, the number of quotations from each of these authors in the *Little Oxford Dictionary of Quotations* (Ratcliffe, 1994) was tallied. For the popular poets the median was 7 entries, for the obscure poets the median was zero. This difference is highly significant (Mann-Whitney test, $p<0.00005$).

Some aspects of the language of the two subsets were then examined. Although the popular poems were on average longer than the obscure ones (median length 155 and 127 words respectively), this difference was not statistically significant (Mann-Whitney test, $p=0.15$). However, a number of significant differences were found: (1) the popular poems had significantly fewer syllables per word in their first lines (Mann-Whitney test, $p=0.035$); (2) popular poems were more likely to begin with an initial line composed entirely of monosyllables (Chi-squared, $p<0.05$); (3) the mean number of letters per word in the popular poems was very significantly less (4.13 versus 4.29) than the obscure poems (unpaired t-test, $p=0.0004$); (4) the vocabulary of the popular poems was on average less rich than that of the obscure ones ($p=0.04$). Syntactic differences were also investigated. Overall a clear tendency for famous poems to use simpler language than obscure poems was found. In poetry, simplicity would seem to be a virtue.

Keywords: Empirical Aesthetics, Quantitative Linguistics, Stylometry, Text Analysis.

1. Background

This paper reports early results from an ongoing empirical study into the stylistic properties of preeminent poems. This project has twin long-term objectives, both still distant. The first is analytical: to formulate analytical rules for telling the difference between successful and unsuccessful poems. The second is synthetic: to formulate heuristics for composing successful rather than unsuccessful poems.

As an initial exploratory exercise, a number of popular English poems were selected and a matching number of obscure or unpopular poems collected as a control group. These were compared in terms of vocabulary, lexical choices and syntax.

Such a Baconian approach (either Roger or Francis Bacon could be taken as a figurehead) might well merit the term "dust-bowl empiricism". It might seem simplistic or even Quixotic to some observers. However, in the present context, it can be justified by reference to earlier work by Martindale (1990), Simonton (1990), Eysenck (1997) and others. These researchers have shown that aesthetic preferences are very far from being purely idiosyncratic. When it comes to aesthetic judgements there is widespread agreement between judges of both sexes, three races, several cultures and many levels of expertise. A simple explanation for such pervasive agreement is that some artworks have intrinsic properties that render them more successful than others.

The present study attempts to seek indicators of such properties in the domain of English verse.

2. Methodological Preliminaries

In order to compare successful with unsuccessful poems, one must have examples of both types. One possible approach is to use an individual poet as his, or her, own control. In other words, to compare the more popular with the less popular works of a single poet. This is the method followed by Simonton (1989; 1990), who took the sonnets of Shakespeare as a single-author case study. He took the number of appearances in 27 anthologies of each sonnet as a direct index of popularity, thus an indirect index of success, and related that to a number of linguistic variables. The great advantage of this procedure is that all extraneous sources of variation due to individual differences between poets are eliminated. Indeed, since Shakespeare probably wrote all 154 sonnets in a period of less than two years, even differences due to the writer's development over time (such as do affect his plays) can be neglected.

The main disadvantage of this strategy is that it cannot uncover factors common to several poets or to a whole tradition of poetry. For that reason the present study concerns itself with 108 authors working over an interval of more than 400 years. It uses an inherently less sensitive design, but, by the same token, any differences that do emerge may be taken as more robust.

Undoubtedly there is a place for both approaches -- single-author studies and multi-author comparisons -- in empirical aesthetics.

3. Selection of Texts

The procedure followed in the present case was first to gather a selection of popular English

poems, then collect an equal number of matching poems as a "control sample". This is analogous to the practice in clinical research of taking a certain number of diseased patients and then comparing them to a control sample of individuals without the disease, each of whom has been matched according to various factors that are thought to be important, such as age and sex. In the jargon of Clinical Trials, this would be called a *retrospective matched case-control study* (Everitt, 1998).

Specifically, the first line of every poem in 20 different general anthologies (published between 1966 and 1997) was typed into a file. Variant spellings were standardized, as was punctuation, and then a program counted the occurrences of each first-line and ranked them by frequency. The first 30 entries in this ranked list are given in Table 1 -- with the authors' names attached. This can be seen as an approximation to a poetic "Top 30".

The number following each author's name and preceding the text of the first line is the frequency count. For example, at the top of this list, Matthew Arnold's "Dover Beach" occurred in 16 of the 20 anthologies, while, at the bottom of this list, Christopher Marlowe's "The Passionate Shepherd to his Love" occurred in nine of them.

Table 1 -- An English Poetic "Top 30"

Arnold, M	16	the sea is calm tonight
Gray, T	15	the curfew tolls the knell of parting day
Coleridge, ST	15	in xanadu did kubla khan
Shakespeare, W	14	shall i compare thee to a summers day?
Blake, W	13	tiger! tiger! burning bright
Shelley, PB	13	i met a traveller from an antique land
Marvell, A	13	had we but world enough and time
Thomas, D	12	now as i was young and easy under the apple boughs
Wordsworth, W	12	earth has not anything to show more fair
Keats, J	11	oh what can ail thee knightatarms
Burns, R	11	oh my loves like a red red rose
Keats, J	11	my heart aches and a drowsy numbness pains
Shakespeare, W	11	let me not to the marriage of true minds
Thomas, E	10	yes i remember adlestrop
Owen, W	10	what passingbells for these who die as cattle?
Byron, GG	10	so well go no more aroving
Keats, J	10	much have i travelled in the realms of gold
Yeats, WB	10	i will arise and go now and go to innisfree
Hopkins, GM	10	glory be to god for dappled things
Shakespeare, W	9	when in disgrace with fortune and mens eyes
Milton, J	9	when i consider how my light is spent
Wyatt, T	9	they flee from me that sometime did me seek
Drayton, M	9	since theres no help come let us kiss and part
Delamare, W	9	is there anybody there? said the traveller
Brooke, R	9	if i should die think only this of me
Wordsworth, W	9	i wandered lonely as a cloud
Hopkins, GM	9	i caught this morning mornings minion king
Browning, EB	9	how do i love thee? let me count the ways
Thomas, D	9	do not go gentle into that good night
Marlowe, C	9	come live with me and be my love

I do not wish to have to justify every aspect of this list. I don't happen to believe, for instance, that Matthew Arnold is the best poet to have written in English, though I do think that "Dover Beach" is a very fine poem. I am quite sure that most readers will find items in this list that they dislike and will note the absence of several favourites of undoubted excellence. My own

favourite poem is also missing from this list. Such discussions, while fascinating, would divert us onto a side track.

For the present purpose it suffices to remind ourselves that, of all the poems ever written, most are completely lacking in poetic merit. The poems listed in Table 1, on the other hand, have appealed to a range of different judges, over many years.

In fact all poems appearing more than five times in this listing were picked, giving 85 poems altogether, by 54 different authors. One of these authors was "Anon" -- the author of the ballad called "Sir Patrick Spens".

The control sample was then formed by finding a matching author for each of the selected authors, then picking the same number of poems by the control author as by the matched popular author. Conditions imposed were that: (1) a control author had to be born within 10 years of the birth date of the matched author; (2) the control author had to be of the same sex as the matched author.

The search for "obscure" authors was entertaining, if rather laborious. It involved, among other things, trawling through second-hand bookshops and more specialized anthologies. The precise details will not be spelled out here; though it must be admitted that the result is not a true random sample. However, a true random sample of all poems, or even all poets, who have written in English is impossible to obtain. The main reason for accepting this less than perfect substitute is the fact that most poems, even those written by well-known poets, are mediocre or worse. Almost any selection process other than picking much-anthologized pieces is bound to lead to a preponderance of mediocre poems simply because the vast majority of published poems are mediocre: they have been made public, have had a chance to become popular favourites, but have remained in obscurity.

Most of the control poets are minor poets. Their works are quite competent. So we are not looking at the difference between outright doggerel and great verse, but at the difference between memorable and forgettable poetry. To call the control poems "flops" is harsh, justified only by the need for a catchy title. Nevertheless, an argument that the 85 control poems are, as a group, better in any meaningful sense than the 85 selected popular poems would be impossible to sustain.

To enable readers to form their own opinions in this matter, Table 2 lists all the poets concerned. The central column, labelled "No.", gives the number of poems selected by both the poets in that row. Thus, for example, William Lisle Bowles, the control for William Blake, contributed 5 poems because five of Blake's appeared in the top 85 -- and so on.

The treatment of "Anon" is slightly exceptional, as no birthdate can be ascertained for either author, and it is not entirely certain that we are dealing with two different authors. In any case, the two anonymous poems were "Sir Patrick Spens" and "Balow". In future studies it may be better to omit anonymous poems altogether.

Table 2 -- Popular Poets and their Controls.

Date	Popular Poet	Q	No.	Date	Obscure Poet	Q
15??	Anon	-	1	15??	Anon	-
1503	Wyatt, T	0	1	1505	Udall, Nicholas	0
1558	Tichborne, C	0	1	1553	Munday, Anthony	0
1563	Drayton, M	2	1	1558	Warner, William	0
1564	Marlowe, C	3	1	1566	Hoskins, John	0
1564	Shakespeare, W	147	5	1563	Sylvester, Joshua	0
1567	Nashe, T	1	1	1566	Bastard, Thomas	1
1572	Donne, J	17	2	1575	Davison, Francis	0
1573	Jonson, B	9	1	1569	Barnes, Barnaby	0
1591	Herrick, R	7	1	1583	Townshend, Aurelian	0
1593	Herbert, G	6	2	1587	Kynaston, Francis	0
1608	Milton, J	36	1	1606	Davenant, William	0
1609	Suckling, J	2	1	1611	Cartwright, William	0
1618	Lovelace, R	2	1	1613	Cleveland, John	0
1621	Marvell, A	4	1	1625	Stanley, Thomas	0
1716	Gray, T	8	1	1721	Akenside, Mark	0
1757	Blake, W	24	5	1762	Bowles, WL	0
1759	Burns, R	13	1	1754	Rowe, Henry	0
1770	Wordsworth, W	21	5	1775	Lloyd, Charles	0
1772	Coleridge, ST	19	2	1775	Lamb, Charles	10
1784	Hunt, JL	0	1	1781	Elliott, Ebenezer	1
1788	Byron, GG	33	2	1785	Peacock, TL	2
1791	Wolfe, C	0	1	1794	Lockhart, JG	0
1792	Shelley, PB	19	2	1802	Praed, Winthrop	0
1793	Clare, J	3	1	1795	Carlyle, Thomas	13
1795	Keats, J	25	5	1796	Coleridge, Hartley	0
1799	Hood, T	2	1	1804	Warburton, Egerton	0
1806	Browning, EB	3	1	1807	Countess Dufferin	0
1809	Tennyson, A	34	1	1811	Thackeray, WM	2

1812	Lear, E	0	1	1819	Jones, Ernest	0
1812	Browning, R	24	3	1817	Bronte, Branwell	0
1822	Arnold, M	20	1	1824	Dobell, Sydney	0
1830	Dickinson, E	2	1	1821	Greenwell, Dora	0
1830	Rossetti, CG	2	1	1829	Siddal, Elizabeth	0
1832	Carroll, L	15	1	1835	Garnett, Richard	0
1840	Hardy, T	9	2	1835	Warren, JL (De Tabley)	0
1844	Hopkins, GM	12	2	1840	Dobson, Austin	3
1859	Housman, AE	6	1	1859	Beeching, HC	0
1865	Kipling, R	22	2	1865	Symons, Arthur	0
1865	Yeats, WB	20	3	1862	Parkes, AJ	0
1873	DelaMare, W	4	1	1873	Ford, FM	0
1874	Frost, R	13	1	1866	Gray, John	0
1878	Thomas, E	1	1	1883	Hulme, TE	0
1878	Masefield, J	1	2	1878	Gibson, Wilfred	0
1885	Lawrence, DH	8	1	1885	Flint, FS	0
1887	Brooke, R	7	1	1892	Aldington, Richard	1
1888	Eliot, TS	38	1	1889	Aiken, Conrad	0
1893	Owen, W	4	2	1897	Sitwell, Sacheverell	0
1902	Smith, S	4	1	1901	Riding, Laura	0
1907	MacNeice, L	6	1	1906	Watkins, Vernon	0
1907	Auden, WH	24	1	1909	Pudney, John	0
1914	Thomas, D	7	2	1913	Forsyth, James	0
1914	Reed, H	3	1	1915	Cave-Browne-Cave, B	0
1922	Larkin, P	11	1	1918	Bell, Martin	0

The column labelled "Q" is the number of quotations in the *Little Oxford Dictionary of Quotations* (Ratcliffe, 1994) by each of the 106 named authors. This information was gathered as a check on the effect of the selection process, and is analysed in the following section.

4. Findings

The results are presented here in four subsections. The first deals with some (partial) checks on the validity of the selection process. The next three deal with characteristic differences between the two groups of poetry in terms of lexical variables, vocabulary, and syntactic features -- i.e. in roughly increasing order of linguistic complexity.

Note that in what follows, unpaired statistical tests have been performed unless otherwise stated. That is to say, the pairing between matched and control authors, which merely served to reduce extraneous sources of variation, was ignored, since it is arbitrary at the level of individual poems, which is the level at which comparisons were made.

Note also that punctuation has been ignored in all calculations in this section.

4.1 Selection Checks

It turned out that the popular poems were longer on average than the obscure ones. The mean number of words in the former category was 246.8 while in the latter it was 194.6. As the distribution of lengths was clearly asymmetrical, a non-parametric (Mann-Whitney) test was performed to assess the significance of this difference. By this test the median lengths (155 and 127 words respectively) were not significantly different ($p=0.1486$, adjusted for ties). In what follows this size difference is therefore ignored.

A check was also performed on the temporal matching between popular poets and their controls. The mean difference between the birthdate of the popular poets and the controls was -0.094 which was not significant by a paired t-test ($t = -0.17$, $p=0.87$). Thus this matching process can be taken as effective.

A curiosity that emerged at this stage was the clearly non-random distribution of birthdates among the 85 popular poets (and their controls, though this latter is artefactual). The most striking feature of this can be observed in Table 2: only one single poet (Thomas Gray, born in 1716) is found in the 136-year period between the birth of Andrew Marvell and that of William Blake. No John Dryden, no Thomas Traherne, no Aphra Benn, no Earl of Rochester, no Samuel Johnson, no Oliver Goldsmith -- not even Alexander Pope. Thus the aversion of our current taste towards the "Augustan" age is clearly illustrated. Whether this affects the validity of the present study would require a separate investigation; at present there is no compelling reason to think that it would.

In addition, an indirect check on the effectiveness of the selection process was performed by looking up the number of quotations attributed to each of the 106 named authors in the *Little Oxford Dictionary of Quotations* (Ratcliffe, 1994). This measure (given in Table 2) is a different index of popularity from that used to select the authors in the first place; but, as Simonton (1989; 1990) has shown, most such indices are highly correlated. For the popular poets the median number of quotations was 7, for the obscure poets the median was zero. This difference is very highly significant (Mann-Whitney test, $p<0.00005$). Table 3 illustrates this contrast from a slightly different viewpoint.

Table 3 -- Quotability of Both Subsamples.

Presence in Little Oxford Dictionary of Quotations :	Quoted at least once	Not quoted at all
Popular Poets	48	5
Obscure Poets	8	45

Such results can be taken as indirect corroboration that the groups differ as intended in terms of authorial impact.

4.2 Size Does Matter: Less is More!

We turn next to low-level variables, such as word-length, measured in syllables and in characters.

Analyzing the initial lines of each poem first, it was found that there were significant differences between the two subgroups. Median characters per word was 3.89 for the popular poems and 4.29 for the obscure poems (Mann-Whitney test, $p=0.0173$). Median number of syllables per word was 1.25 for the popular poems and 1.2857 for the obscure ones (Mann-Whitney, $p=0.0353$). Popular poems were also more likely to start with a line composed entirely of monosyllables, as shown by Table 4.

Table 4 -- Frequency of Monosyllabic Opening Lines.

Syllables in first line:	all words monosyllabic	some words polysyllabic
Popular poems	17	68
Obscure poems	7	78

A Chi-squared test on this table gave a Chi-squared value of 3.9298 (with 1 d.f.) after applying Yates's correction. This value is significant at the $p<0.05$ level. Thus the initial lines of popular poems tend to employ shorter words than those of obscure poems.

Taking the poems as whole, a similar pattern emerges. For both types of poem the average number of characters and of syllables per word is summarized in Table 5.

Table 5 -- Average Word Lengths in Characters & Syllables.

Average word lengths for whole poems :	Mean / Characters	Median / Characters	Mean / Syllables	Median / Syllables
Popular poems	4.1323	4.1484	1.2816	1.2789
Obscure poems	4.2944	4.3087	1.3263	1.3364

Since these data were approximately normally distributed, (unpaired) t-tests were performed. With characters, t = -3.58, p = 0.0004; with syllables, t = -3.22, p = 0.0015. On either measure, therefore, there is a highly significant difference. (Non-parametric, Mann-Whitney, tests gave almost identical results.)

Popular poems use shorter words than obscure ones.

4.3 Word Frequency and Vocabulary Richness

It has been known since the work of Zipf (1935) that there is a systematic tendency for commonly used words to be shorter than less commonly used words in any language. So a study was also made of the relative frequency of the words used in all the poems, in the following manner.

Hofland & Johansson (1982) give a listing of all word forms used at least 10 times in either the LOB (Lancaster-Olso-Bergen) corpus of British English or the Brown Corpus of American English (Francis & Kucera, 1982). Both these corpora were collated from prose written in 1961 and both consist of approximately a million word-tokens. This list of words, 9175 in total, has been entered into a file by the present author and the total number of occurrences in both corpora (Brown+LOB) aggregated. A program was written to read through the 170 poems in our sample and look up each word in this joint dictionary. If it was not found, it was given a default frequency of nine -- one less than the least frequently occurring word.

As this distribution is extremely skewed ("the" having a total frequency of 138,285 out of approximately 2 million) the base-10 logarithm of each frequency was taken. Then, for each poem, the mean logged frequency was computed. Although this data is derived from prose, and although it is based on evidence near the end of the 400-year period under investigation, it does provide an objective index of the commonness of words in the English language.

The medians of these mean logged frequency scores for the popular and obscure poems were 3.0407 and 2.9064 respectively. A Mann-Whitney test showed this difference to be very highly significant (p = 0.0009). Thus the words in the popular poems tend to be more common than those in the obscure poems.

The percentage of words not found in this LOB/Brown dictionary of 9175 entries was also recorded, which provides an alternative index of rare-word usage. The median percentage of unfound words in the 85 popular poems was 12.308, while that in the 85 obscure poems was 15.957. A Mann-Whitney test showed this difference also to be very highly significant (p = 0.0006). As this variable had only a small non-significant correlation (r = -0.057) with the serial position of the poems (an approximation to date of composition), it can be taken as an index of rarity rather than recency.

Hence there is evidence that popular poems use more common words and fewer rare words than obscure poems.

High rates of usage of rare words are often associated with relatively rich vocabularies, so it was decided to compare the two sets of poems for vocabulary richness.

Many different measures of vocabulary richness have been proposed and used, following the

work of Yule (1944), Herdan (1966), Brainerd (1972) and others. See, for example, Holmes (1985). One of the simplest is the Bilogarithmic type-token ratio, Log(V)/Log(N), where V is the number of distinct types in a text, i.e. vocabulary size, and N is the number of tokens. Some writers (e.g. McKinnon & Webster, 1971; Leavitt & Mitchell, 1977) have even suggested that this ratio is unaffected by text size, though in fact it is, as can be seen from Figure 1.

[Figure 1 about here.]

Figure 1 shows this variable Log(V)/Log(N) (or BiLogTTR) plotted against Log(N) for the 170 poems in our sample. Logs to the base 10 were used. The correlation between BiLogTTR and Log(N) of r = -0.461 is significant. So this variable is related to text length. However this linear relationship can be removed by regressing BiLogTTR on Log(N) and taking the **residuals** or deviations from this regression formula, below,

$$\text{BiLogTTR} = 1.01 - 0.0448 * \text{Log(N)}$$

as a measure of vocabulary richness. The regression line of this equation is also plotted in Figure 1.

When this is done, a significant difference is found: the median residual for the popular poems is 0.083, while for the obscure poems it is 0.2462 (p = 0.0387, Mann-Whitney test).

This shows that the obscure poems tend to have a richer vocabulary than the popular ones.

4.4 Usage of Frequent Words and Syntactic Tags

While several statistically significant differences between the popular and obscure poems have been reported above, none of the individual variables investigated so far would serve very well as a discriminator between the two categories on its own. It was therefore decided to try a multivariate approach to discriminating between these two categories.

Following the pioneering work of Mosteller & Wallace (1964/1984), many researchers have used high-frequency words as the basis for discriminating literary texts on the basis of genre or authorship. See, for example: Burrows (1989; 1992), Craig (1992), Binongo (1994) and Holmes & Forsyth (1995).

Recently, however, Baayen et al. (1996) have argued that high-frequency words, which are mostly function words, act in this context as surrogate indicators of syntactic constructions and hence that studies of this kind would do better, where possible, to look at syntactic habits more directly.

To investigate this question, two linear discriminant analyses were performed to distinguish between the two classes of poems in our 170-item sample. The first used the most frequently occurring 40 words in the joint sample of 170 poems as features; the second used the most frequently occurring 40 syntactic tags in the joint sample as features.

Table 6 -- The 40 Commonest Words in the 170 Poems.

Word	Frequency	Rank	Percent	Cum%
the	2248	1	5.9379	5.9379
and	1599	2	4.2236	10.161
of	938	3	2.4776	12.639
a	795	4	2.0999	14.739
to	640	5	1.6905	16.429
i	611	6	1.6139	18.043
in	582	7	1.5373	19.581
that	395	8	1.0433	20.624
with	374	9	0.9879	21.612
my	370	10	0.9773	22.589
is	294	11	0.7765	23.366
his	256	12	0.6762	24.042
for	250	13	0.6603	24.702
on	241	14	0.6365	25.339
but	238	15	0.6286	25.968
not	238	16	0.6286	26.596
as	233	17	0.6154	27.212
it	228	18	0.6022	27.814
all	227	19	0.5996	28.414
was	209	20	0.552	28.966
from	179	21	0.4728	29.439
her	177	22	0.4675	29.906
no	169	23	0.4464	30.353
or	168	24	0.4437	30.796
at	168	25	0.4437	31.240
be	157	26	0.4147	31.655
we	153	27	0.4041	32.059
by	148	28	0.3909	32.450
thy	147	29	0.3882	32.838
you	145	30	0.383	33.221
me	140	31	0.3698	33.591
have	138	32	0.3645	33.955
he	138	33	0.3645	34.320
this	135	34	0.3566	34.677
when	129	35	0.3407	35.017
so	127	36	0.3354	35.353
they	124	37	0.3275	35.680
are	123	38	0.3249	36.005
their	117	39	0.309	36.314
love	117	40	0.309	36.623

To obtain syntactic information the texts were sent by email to the Birmingham University tagger. This is a free service which can be used by sending an electronic mail message to

tagger@clg.bham.ac.uk

which takes a plain ASCII file of English-language text and appends part-of-speech tags to each word. The tagset used, of about 60 different tags, is a slightly modified version of the Brown tagset (Francis & Kucera, 1982). It includes ??? for words that the tagger fails to identify. This may not be the best tagging software ever developed, but it is free, fast, and its tagset is not hugely elaborate.

As illustration, an extract from a poem tagged by the Birmingham tagger is listed below as Table 7. Note that the vertical format, with one token per line, is produced by the tagger: input texts were normal running text with line breaks as on a printed page.

Table 7 -- Sample of Tagged Text.

```
[               (       [
FIDELE          NP      fidele
'S              POS     's
DIRGE           NN      dirge
]               )       ]
FEAR            NN      fear
no              DT      no
more            JJR     more
the             DT      the
heat            NN      heat
o               ???     o
'               '       '
the             DT      the
sun             NN      sun
,               ,       ,
Nor             CC      nor
the             DT      the
furious         JJ      furious
winter          NN      winter
's              POS     's
rages           VBZ     rage
;               :       ;
Thou            PP      thou
thy             PP$     thy
worldly         JJ      world
task            NN      task
hast            VBP     hast
done            DON     do
,               ,       ,
Home            NN      home
art             NN      art
gone            VBN     go
,               ,       ,
and             CC      and
ta'en           ???     ta'en
thy             PP$     thy
wages           NNS     wage
;               :       ;
Golden          NP      golden
lads            NNS     lad
and             CC      and
girls           NNS     girl
all             DT      all
must            MD      must
As              IN      as
chimney-sweepers        NNS     chimney-sweepers
,               ,       ,
come            VB      come
to              TO      to
dust            NN      dust
.               .       .
```

Using both types of information (frequent words and syntactic tags) a stepwise discriminant function was performed. Thus the statistical package (SPSS) used a heuristic method to pick from the 40 variables available the most discriminatory subset. Then these variables were used in another package (Minitab, because it allows cross-validation) to derive a linear discriminant function for classifying each poem. These functions, derived from the full data set of 170 records, were recorded. The classification success rate, using the leave-1-out method of cross-validation, was also recorded.

Using frequent words as variables this procedure picked just two variables, "and" and "I", both more frequent in the popular than obscure poems. The standardized distance between the two groups was 0.6204 and the cross-validated predictive success rate was 64.1%. Standardized canonical discriminant function scores are given below.

and	0.8276
I	0.5193

Using syntactic tags as variables, this procedure selected five variables, which are listed together with their standardized canonical discriminant function scores in Table 8.

Table 8 -- Discriminating Syntactic Tags.

Tag	**Score**	**Part of Speech**	**Examples**
NN	-0.4621	Noun, singular	love, heart, man, day
DT	0.4235	Determiner	the, a, all, no
PP	0.5758	Personal Pronoun	I, it, we, you
CC	0.5891	Coordinating Conjunction	and, but, or, nor
VBP	-0.3800	Verb, present tense	breathe, seem, think, hope

The standardized distance between the two groups was 1.0044 and the cross-validated predictive success rate was 66.5%. The difference between these two success rates is in the direction hypothesized by Baayen et al. (1996), although it is marginal.

Thus the first analysis has found that the coordinating conjunction "and" and the personal pronoun "I" are both more frequent in the popular than the obscure poems. The second analysis has found that nouns (NN) and the present (not third-person singular) forms of verbs (VBP) are more frequent in the obscure poems, while determiners (DT), personal pronouns (PP) and coordinating conjunctions (CC) in general are more frequent in the popular poems.

5. Discussion

The present study has found that:

(1) popular poems tend to use shorter words, whether measured by syllables or by character per word, than obscure poems;

(2) obscure poems contain a significantly higher proportion of rare words than popular ones;

(3) obscure poems tend to employ a more diverse vocabulary than popular ones;

(4) popular poems exhibit a high rate of coordinating conjunctions (especially the word "and") and of personal pronouns (especially "I") compared with obscure ones;

(5) obscure poems tend to have a higher rate of singular nouns and present-tense verbs than popular poems.

Could we characterize these differences in a single sentence? One possible summing-up would be: the language of the popular poems is basic, functional, person-centred (indeed self-centred) and somewhat repetitive, compared with that of the obscure poems. On the whole, these attributes are distinctive of spoken, as opposed to written, language.

Still, using the information analyzed to discriminate between popular and obscure poems gives an error rate of more than one in three; so there is plenty more work to be done. One line of future enquiry will be to look at tag-transitions rather than simple tag frequencies or rates -- giving at least some information on the essentially serial nature of syntactic structure. Another will be to look at semantic information as well as lexical and syntactic, following the lead of Martindale (1990) -- though none of the readily available content-analytic resources is particularly well-suited to the task in hand.

5.1 Relation to Previous Work

As noted above, some of these findings have a bearing on results reported by previous authors. The slight superiority of the linear discriminant function based on syntactic tags to that based on frequent words (the same number of features in both cases) tends to support the proposition of Baayen et al. (1996) that frequent words as discriminators are merely surrogates for syntactic information that could more effectively be tapped directly. The additional evidence provided by the present study is by no means conclusive, but it does suggest that the effort of getting a more accurate tagger or correcting the tagged texts by hand might be worthwhile.

The results on vocabulary richness in section 4.4 contradict a conclusion drawn by Simonton (1989) in his study of Shakespeare's sonnets:

> "The better sonnets are distinguished by a higher type-token ratio" (Simonton, 1989, p. 710).

However, before we conclude that Shakespeare's sonnets are an exceptional case, we should

note that in a follow-up study, which analyzed each sonnet in four segments (three quatrains and a final couplet), Simonton found a more complex picture, and concluded that:

> "Shakespeare was unlikely to resort to seldom-used words when conceiving the concluding six lines of his best sonnets." (Simonton, 1990, p. 261)

Nevertheless, there is a degree of conflict on this point, and it may well be necessary to perform some other single-author studies to understand better the complex relationship between vocabulary richness and poetic merit.

5.2 Concluding Remarks

It could be objected that the five specific findings listed at the head of this section relate to rather superficial linguistic attributes, far removed from whatever it is that "breathes fire" into poetry. It might also be objected that they seem obvious.

However, as far as the first objection is concerned, it should be noted that this is a field, the study of poetic merit, which has long suffered from unconstrained theorizing. Even results about low-level linguistic features, provided that they are objective, help to constrain the wilder excesses of our thinking and thereby help us refine our theories.

Regarding the second point, it is a fact of human psychology that most results seem obvious -- with hindsight. Moreover, at least one of the findings reported here is counter-intuitive, namely that obscure poems tend to employ a richer vocabulary than popular ones.

In addition, once we have identified objective correlates of poetic success, we can move further towards the long-term goal of the present work within an experimental framework. For example, it is planned to take translations from poems in languages little read in England (e.g. Chuvash) and manipulate various lexical and syntactic features such as vocabulary richness or mean word-length to create variant versions of original poems that differ on different dimensions. Then these can be presented to readers to be ranked or rated and the effects of the experimental manipulations on readers' preferences assessed.

Acknowledgements

I have received absolutely no financial support from any academic funding agency for the present work. Presumably such agencies consider it completely pointless, which is probably a good sign.

On the other hand, I would like to thank Oliver Mason and the corpus research group at Birmingham University for making their tagging software freely available. For details see:

> http://www-clg.bham.ac.uk

I also thank my home institution, UWE Bristol, for various facilities used during this research.

References

Baayen, H., van Halteren, H. & Tweedie, F.J. (1996). Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution. *Literary & Linguistic Computing*, 11(3), 121-131.

Binongo, J.N.G. (1994). Joaquin's Joaquinesquerie, Joaquinesquerie's Joaquin: a Statistical Expression of a Filipino Writer's Style. *Literary & Linguistic Computing*, 9(4), 267-279.

Brainerd, B. (1972). On the Relation between Types and Tokens in Literary text. *J. Appl. Prob.*, 9, 507-518.

Burrows, J.F. (1989). "An Ocean where each Kind ...": Statistical Analysis and some Major Determinants of Literary Style. *Computers & the Humanities*, 23, 309-321.

Burrows, J.F. (1992). Not Unless You Ask Nicely: the Interpretive Nexus between Analysis and Information. *Literary & Linguistic Computing*, 7(2), 91-109.

Craig, D.H. (1992). Authorial Styles and Frequencies of Very Common Words: Jonson, Shakespeare and the Additions to The Spanish Tragedy. *Style*, 26, 199-220.

Everitt, B.S. (1998). *The Cambridge Dictionary of Statistics*. Cambridge University Press, Cambridge.

Eysenck, H.J. (1997). The Objectivity and Lawfulness of Aesthetic Judgements. In: Leonid Dorfman et al. (eds.) *Emotion, Creativity and Art*. Perm State Institute of Arts & Culture, Perm, Russia.

Francis, W.N. & Kucera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, Boston.

Herdan, G. (1966). *The Advanced Theory of Language as Choice and Chance*. Springer-Verlag, New York.

Hofland, K. & Johansson, S. (1982). *Word Frequencies in British and American English*. Norwegian Computing Centre for the Humanities, Bergen.

Holmes, D.I. (1985). The Analysis of Literary Style: a Review. *J. Royal Statistical Soc. (A)*, 148(4), 328-341.

Holmes, D.I. & Forsyth, R.S. (1995). The "Federalist" Revisited: New Directions in Authorship Attribution. *Literary & Linguistic Computing*, 10(2), 111-127.

Leavitt, J.A. & Mitchell, J.L. (1977). SPAN: a Lexicostatistical Measure and Some Applications. In: S. Lusignan & J.S. North (eds.) *Computing in the Humanities*. Univ. Waterloo Press, Waterloo, Ontario.

Martindale, C. (1990). *The Clockwork Muse*. Basic Books, N.Y.

McKinnon, A. & Webster, R. (1971). A Method of "author" Identification. In: R.A. Wisbey (ed.) *The Computer in Literary and Linguistic Research*. Cambridge Univ. Press.

Mosteller, F. & Wallace, D.L. (1984). *Applied Bayesian and Classical Inference: the Case of The Federalist Papers*. Springer-Verlag, New York. [First edition published 1964.]

Ratcliffe, S. (1994) ed. *The Little Oxford Dictionary of Quotations*. Oxford University Press, Oxford.

Simonton, D.K. (1989). Shakespeare's Sonnets: a Case of and for Single-Case Historiometry. *J. of Personality*, 57(3), 695-721.

Simonton, D.K. (1990). Lexical Choices and Aesthetic Success: a Computer Content Analysis of 154 Shakespeare Sonnets. *Computers & the Humanities*, 24, 251-264.

Yule, G.U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge Univ. Press.

Zipf, G. (1935). *The Psycho-Biology of Language*. Houghton-Mifflin, Boston.

Appendix 1

Readers may be interested to know about another tagger, the TOSCA tagger from Nijmegen, which is also free. It can be obtained from

ftp://lands.let.kun.nl/pub/tosca/tlbtag

though I personally haven't yet had time to compare it with the Birmingham tagger, except to note that it uses a richer set of tags, and that it doesn't produce a lemmatized list of the input words.

Appendix 2

The foregoing discussion has concentrated on ranking poems by popularity rather than poets, which is not quite the same thing. In fact the latter is less difficult. Readers may find it entertaining to view the following list, based on a larger sample of anthologies than that used to rank the most popular poems (33 rather than 20). It shows the first 32 poets, treating Anon as an individual.

The integer part of each poet's score simply counts how many anthologies that poet appeared in. The fractional part is a tie-breaker based on the number of entries within each anthology.

**** POET : Major English-language Poets, in order :**

1	William Shakespeare	32.8465413
2	Alfred Tennyson	32.6537434
3	Robert Browning	32.4947161
4	Percy B. Shelley	32.4840332
5	William Wordsworth	31.6833568
6	William B. Yeats	31.5931871
7	John Keats	31.5921994
8	Dylan M. Thomas	31.4103728
9	William Blake	30.4998407
10	George G. Byron	30.4689275
11	Samuel T. Coleridge	30.4561009
12	Gerard M. Hopkins	30.4427786
13	Matthew Arnold	30.3753911
14	Christina G. Rossetti	30.3066367
15	Wystan H. Auden	29.3999652
16	Alfred E. Housman	29.293376
17	Thomas S. Eliot	28.4248351
18	Thomas Hardy	28.395584
19	Andrew Marvell	28.3243535
20	Wilfred Owen	28.2519966
21	John Milton	27.5418018
22	Robert Burns	27.3842323
23	John Donne	26.5319061
24	Philip Larkin	26.3167587
25	Louis MacNeice	26.2723867
26	George Herbert	26.2704662
27	Emily Dickinson	25.3387563
28	Robert Frost	25.2963982
29	Robert Herrick	25.27709
30	Thomas Gray	25.2715932
31	Edward Thomas	25.2122471
32	Anon	24.2943465

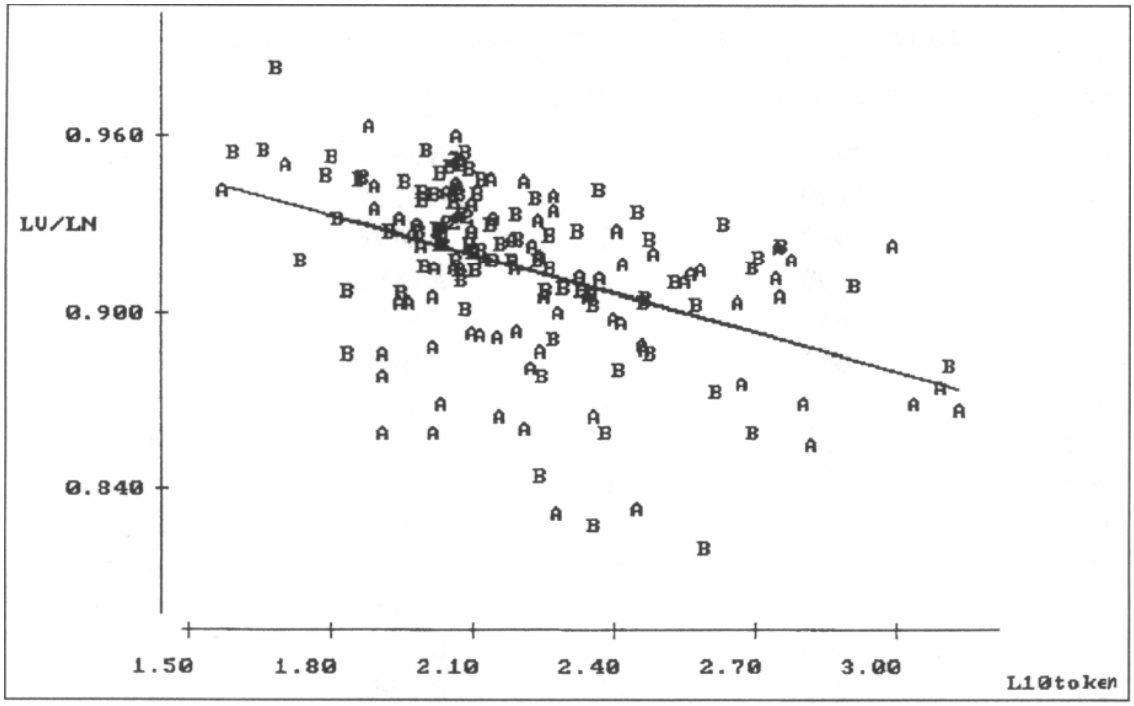

Figure 1.