'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship[1]

John Burrows
University of Newcastle, Australia

Abstract

This paper is a companion to my 'Questions of authorship: attribution and beyond', in which I sketched a new way of using the relative frequencies of the very common words for comparing written texts and testing their likely authorship. The main emphasis of that paper was not on the new procedure but on the broader consequences of our increasing sophistication in making such comparisons and the increasing (although never absolute) reliability of our inferences about authorship. My present objects, accordingly, are to give a more complete account of the procedure itself; to report the outcome of an extensive set of trials; and to consider the strengths and limitations of the new procedure. The procedure offers a simple but comparatively accurate addition to our current methods of distinguishing the most likely author of texts exceeding about 1,500 words in length. It is of even greater value as a method of reducing the field of likely candidates for texts of as little as 100 words in length. Not unexpectedly, it works least well with texts of a genre uncharacteristic of their author and, in one case, with texts far separated in time across a long literary career. Its possible use for other classificatory tasks has not yet been investigated.

Correspondence:
John Burrows,
72 Atherton Close,
Rankin Park,
NSW 2287, Australia.
E-mail:
john.burrows@netcentral.com.au

1 It was originally presented as the Roberto Busa Award Lecture for 2001 at the ACH–ALLC Conference at New York University. It is to be published in the conference-issue of *Computers and the Humanities* (Burrows, 2003).

The point of departure for the development of the 'Delta procedure' (as I call it) was the observation that the methods of comparison and authorial attribution currently employed in computational stylistics are better fitted for 'closed games' than for more open ones. The closed games take two forms. Where only two or three writers are eligible candidates for the authorship of a particular text and where that text is of a sufficient length, we are now well equipped to form strong inferences about their rival claims. The classic study of this kind is Mosteller and Wallace (1964). Holmes (2001) offers an excellent recent specimen. Where the real question is whether or not a particular writer (and no other) is the author, we are equally well equipped to test his or her claims. Tweedie *et al.* (1998) and Burrows and Craig (2001) offer recent specimens of this kind. But in 'open games', where we are faced with an anonymous text but have little or no outside evidence to identify the most likely candi-

dates, our current methods must be employed in an exhaustive and possibly fruitless series of iterations.

A reliable means of detecting unique authorial fingerprints (of whose very existence we do not yet have either proof or promise) would clearly be the best way of resolving these open games. But, in its absence, there is room for a simple measure capable of distinguishing the most likely candidates from a large group and also, where no candidate lays a sufficient claim, of indicating that it might be wise to look further afield. For want of such a measure, we are still bound by Bailey's dictum (1979, p. 7), proposed over 20 years ago and lately put even more strictly by Binongo and Smith (1999, p. 464). We should confine ourselves, they hold, to cases where the choice lies within a narrow range of well-matched sets and we should proceed with only two authors' texts at a time. But, at least in the initial stages of an inquiry, the 'Delta procedure' allows us to shake off these constraints. After employing it to identify the strongest candidates, we can use our current methods to choose among them. The open game is thus transformed into a closed game.

Most of the methods currently employed in computational stylistics rest upon multivariate statistical comparisons between some characteristics of a given specimen and those of an appropriate set of norms. The characteristics, which are used as statistical variables, comprise the relative frequencies of various simple phenomena such as alphabetic characters, strings of characters, whole words, or common grammatical forms. Each of these classes of variables has its advantages and disadvantages, and each has its adherents among scholars in the field. Forsyth and Holmes (1996) offer a reasoned overview. The advantage of working with whole words rests on their accessibility and their meaningfulness. They help us, in particular, to form close and fruitful inferences about the outcome of an inquiry. Whichever class of variables is chosen, it has become customary, in recent years, to allow the particular variables to 'declare themselves', thus obviating, as far as possible, the danger of a predetermined outcome. The words used, for example, might be the 100 most common in the database that provides the norms for a particular inquiry. In this sort of work on language, so our researches teach us, a wealth of variables, many of which may be weak discriminators, almost always offer more tenable results than a smaller number of strong ones. Strong features, perhaps, are easily recognized and modified by an author and just as easily adopted by disciples and imitators. At all events, a distinctive 'stylistic signature' is usually made up of many tiny strokes.

The multivariate statistical instruments now most used in computational stylistics are designed to elicit subtle trends in complex sets of figures. The results obtained from principal component analysis, for example, are usually rendered in bi-axial or tri-axial graphs where the pattern of the entries allows far-reaching inferences to be drawn. Whenever a specimen is added or removed, the whole pattern alters in a fashion that does not admit strict comparisons between graph and graph. It is as if the ingredients of a mixture were being altered and a fresh state studied each time.

Motivation

Was nehmen
als Features?
- wie wurde es vorher
gemacht?
- was sollte nun
genommen werden?

In experienced hands, such methods yield excellent results. But they are obviously unsuitable for the crude but useful task of ranking many candidates in a single, all-embracing hierarchy, thus singling out the statistically most eligible among them. Even a ranked series of aggregates or means, I told myself, might serve that purpose if a sound basis were available. The path forward from this point was long obscured by the fact that, because the scores for any given specimen on the chosen set of variables diverge in both directions from the norms for the database, an aggregate or mean divergence would comprise an arbitrary mixture of positives and negatives.

Although the differences between positives and negatives—high scores, say, for *the* in this specimen but low ones for *I* and *me*—are most instructive, they are not the heart of the matter. An expression of difference, pure difference, is what we seek. If all the positive and negative divergences were rendered as absolute divergences, their overall aggregate or their mean might be of interest.[2] A 'delta-score' is just such a mean divergence. The term 'Delta' (best rendered when possible as 'Δ') was chosen to represent D for Difference and also as a gesture of respect for those heroic pioneers in our field who worked without benefit of computers. Among their various attempts to derive simple expressions of stylistic difference, Udney Yule's Characteristic K was the most fruitful.

= mittlere Abweichung

The first step in the procedure is to establish a frequency-hierarchy for the most common words in a large group of suitable texts. The texts are arranged in subsets representing the work of numerous authors appropriate to the task in hand. With texts of a bygone era, it is usual and desirable to standardize spelling and to expand contracted forms of expression to reduce the influence of trivial or accidental variations. (Just such variations were studied by some of the pioneers of stylometry. But when one works with common word-counts, they are merely a distortion.) It has also been our practice, in Newcastle, to tag some of the more common homographic forms to distinguish the different uses of words such as *so* and *that*. When the word-counts have been made, the frequencies are standardized as proportions of each authorial subset so that the larger subsets do not exert an undue influence on the composition or ranking of the hierarchy.

Vorbemerkung zu Textaufbau

Database

Working on these lines, we formed a database of verse by twenty-five poets of the English Restoration period.[3] These yielded the frequency-hierarchies used for several recent studies of authorship based on principal component analysis. They also yielded the norms for the Delta project. For this project, I have added a further range of texts, all independent of the main set and all of unquestioned if not unquestionable authorship. (It is impossible to be confident of the authorship of every member of a large, mixed set of Restoration poems. But, having gone to reputable sources, I shall stand by the results.)

Table 1, based on a small Microsoft Excel worksheet, offers a 'closed version' of the procedure, bringing the top thirty words of the main set of Restoration verse to bear on a simple question. Can we demonstrate that John Milton has a better claim to a selection from *Paradise Lost*—27,154

2 Experiments with overall medians yielded less accurate results than those to be described.

3 The present corpus of 540,244 words ranges widely across the work of the following twenty-five poets: Aphra Behn (1640–89) 21,705 words; Alexander Brome (1620–66) 29,539; Samuel Butler (1612–80) 30,932; William Congreve (1670–1729) 30,917; Charles Cotton (1630–87) 12,625; Abraham Cowley (1618–67) 19,272; Sir John Denham (1615–69) 30,092; Charles Sackville, Earl of Dorset (1638–1706) 9,586; John Dryden (1631–1700) 18,238; Thomas D'Urfey (1653–1723) 18,757; Robert Gould (1660?–1709?) 29,110; Andrew Marvell (1621–78) 23,282; John Milton (1608–74) 18,924; John Oldham (1653–83) 32,462; Katherine Phillips (1631–64) 29,004; Matthew Prior (1664–1721) 32,000; Alexander Radcliffe (floruit 1669–96) 11,889; John Wilmot, Earl of Rochester (1648–80) 12,725; Sir Charles Sedley (1639?–1701) 10,304; Elkanah Settle (1648–1724) 24,080; Thomas Shadwell (1642?–92) 14,540; Jonathan Swift (1667–1745) 30,974; Nahum Tate (1652–1715) 20,333; Edmund Waller (1606–87) 16,443; Anne Wharton (1659–85) 12,511. Most of the corpus was prepared by John Burrows and Harold Love, assisted by Alexis Antonia and Meredith Sherlock. The Marvell subset was contributed by Christopher Wortham.

Table 1 Specimen of procedure

	A	B	C	D	F	G	I	J	K	L	N	O	P	Q	S	T	U	V	W	X	Y	Z
1			**Main set**		**Milton**		***Paradise Lost***				***World's Infancy***				***Paradise Regained***				***Samson Agonistes***			
2							**count**			**30**	**count**			**30**	**count**			**30**	**count**			**30**
3							**sum**			**31.489**	**sum**			**36.164**	**sum**			**32.247**	**sum**			**33.814**
4							**mean (= "delta")**			**1.050**	**mean (= "delta")**			**1.205**	**mean (= "delta")**			**1.075**	**mean (= "delta")**			**1.127**
5							**stdev**			**0.770**	**stdev**			**1.163**	**stdev**			**1.227**	**stdev**			**1.087**
6			**Mean**	**Stdev**	**Scores**	**z-scores**	**Scores**	**z-scores**	**Diff.**	**Abs. diff.**	**Scores**	**z-scores**	**Diff.**	**Abs. diff.**	**Scores**	**z-scores**	**Diff.**	**Abs. diff.**	**Scores**	**z-scores**	**Diff.**	**Abs. diff.**
7	1	the	4.242	0.630	4.719	0.757	4.091	–0.239	–0.996	0.996	7.866	5.753	4.996	4.996	3.619	–0.988	–1.746	1.746	2.809	–2.274	–3.031	3.031
8	2	and	3.770	0.501	4.407	1.272	4.165	0.789	–0.483	0.483	3.474	–0.590	–1.862	1.862	4.441	1.340	0.068	0.068	3.298	–0.940	–2.212	2.212
9	3	of	1.821	0.315	2.420	1.905	2.769	3.015	1.110	1.110	2.169	1.106	–0.799	0.799	2.765	3.002	1.097	1.097	2.561	2.353	0.448	0.448
10	4	a	1.601	0.430	0.893	–1.645	0.696	–2.103	–0.458	0.458	1.296	–0.708	0.936	0.936	0.873	–1.691	–0.047	0.047	1.094	–1.177	0.468	0.468
11	5	to(i)	1.419	0.272	1.247	–0.634	1.289	–0.480	0.154	0.154	0.918	–1.846	–1.212	1.212	1.389	–0.111	0.523	0.523	1.824	1.491	2.124	2.124
12	6	in(p)	1.358	0.189	1.554	1.035	1.720	1.916	0.881	0.881	1.476	0.624	–0.411	0.411	1.536	0.940	–0.095	0.095	1.552	1.028	–0.007	0.007
13	7	his	1.154	0.323	1.062	–0.284	1.532	1.171	1.454	1.454	1.359	0.635	0.919	0.919	1.287	0.413	0.696	0.696	1.009	–0.448	–0.165	0.165
14	8	with	1.022	0.208	1.480	2.202	1.484	2.224	0.022	0.022	0.972	–0.239	–2.441	2.441	1.141	0.572	–1.630	1.630	1.436	1.991	–0.211	0.211
15	9	to(p)	1.014	0.131	0.999	–0.119	1.245	1.761	1.880	1.880	0.819	–1.493	–1.373	1.373	1.663	4.957	5.077	5.077	1.428	3.161	3.281	3.281
16	10	is	0.938	0.312	0.502	–1.397	0.239	–2.238	–0.841	0.841	1.233	0.944	2.341	2.341	0.465	–1.515	–0.118	0.118	0.442	–1.588	–0.191	0.191
17	11	but	0.923	0.195	0.676	–1.268	0.696	–1.167	0.101	0.101	0.378	–2.801	–1.533	1.533	0.765	–0.814	0.453	0.453	0.916	–0.038	1.230	1.230
18	12	he	0.803	0.241	0.465	–1.403	0.703	–0.413	0.990	0.990	0.603	–0.830	0.573	0.573	0.784	–0.079	1.324	1.324	0.435	–1.529	–0.126	0.126
19	13	all	0.781	0.193	0.518	–1.366	0.836	0.283	1.649	1.649	0.720	–0.318	1.048	1.048	0.975	1.003	2.369	2.369	0.830	0.254	1.620	1.620
20	14	I	0.766	0.391	0.882	0.297	0.700	–0.171	–0.467	0.467	0.711	–0.142	–0.438	0.438	1.198	1.103	0.806	0.806	1.676	2.326	2.030	2.030
21	15	it	0.766	0.239	0.386	–1.591	0.151	–2.575	–0.984	0.984	0.558	–0.870	0.722	0.722	0.299	–1.953	–0.361	0.361	0.450	–1.322	0.270	0.270
22	16	as	0.710	0.224	0.618	–0.410	0.737	0.119	0.529	0.529	0.540	–0.760	–0.350	0.350	0.701	–0.041	0.369	0.369	0.722	0.053	0.463	0.463
23	17	their	0.641	0.237	0.513	–0.540	0.795	0.653	1.193	1.193	0.432	–0.880	–0.340	0.340	0.522	–0.498	0.042	0.042	0.761	0.506	1.046	1.046
24	18	her	0.623	0.336	0.851	0.678	0.435	–0.560	–1.237	1.237	0.756	0.396	–0.282	0.282	0.312	–0.923	–1.601	1.601	0.287	–0.998	–1.675	1.675
25	19	not	0.616	0.174	0.592	–0.138	0.847	1.324	1.462	1.462	0.432	–1.054	–0.916	0.916	0.841	1.290	1.428	1.428	1.180	3.231	3.369	3.369
26	20	be	0.586	0.167	0.555	–0.187	0.401	–1.109	–0.921	0.921	0.459	–0.763	–0.576	0.576	0.503	–0.496	–0.309	0.309	0.520	–0.397	–0.209	0.209
27	21	you	0.580	0.252	0.174	–1.608	0.037	–2.154	–0.546	0.546	0.261	–1.265	0.344	0.344	0.006	–2.275	–0.666	0.666	0.023	–2.208	–0.599	0.599
28	22	they	0.564	0.234	0.270	–1.259	0.464	–0.428	0.830	0.830	0.396	–0.719	0.540	0.540	0.370	–0.831	0.427	0.427	0.310	–1.084	0.175	0.175
29	23	for(p)	0.559	0.114	0.270	–2.539	0.000	–4.905	–2.366	2.366	0.342	–1.903	0.637	0.637	0.280	–2.444	0.095	0.095	0.466	–0.817	1.722	1.722
30	24	by(p)	0.555	0.106	0.412	–1.349	0.689	1.260	2.608	2.608	0.432	–1.162	0.187	0.187	0.822	2.518	3.866	3.866	0.582	0.254	1.603	1.603
31	25	my	0.512	0.370	0.587	0.201	0.258	–0.687	–0.888	0.888	0.351	–0.435	–0.636	0.636	0.472	–0.110	–0.311	0.311	1.226	1.928	1.727	1.727
32	26	we	0.510	0.275	0.159	–1.279	0.265	–0.891	0.388	0.388	0.468	–0.153	1.126	1.126	0.127	–1.392	–0.113	0.113	0.124	–1.404	–0.125	0.125
33	27	from	0.500	0.127	0.534	0.265	0.884	3.019	2.754	2.754	0.567	0.527	0.262	0.262	0.771	2.132	1.866	1.866	0.520	0.157	–0.108	0.108
34	28	that(rp)	0.476	0.228	0.925	1.964	0.313	–0.715	–2.680	2.680	0.234	–1.061	–3.026	3.026	0.172	–1.333	–3.297	3.297	0.217	–1.135	–3.099	3.099
35	29	or	0.471	0.165	0.856	2.333	0.906	2.636	0.302	0.302	0.153	–1.929	–4.263	4.263	1.064	3.595	1.261	1.261	0.908	2.648	0.315	0.315
36	30	our	0.460	0.268	0.270	–0.711	0.354	–0.397	0.314	0.314	0.558	0.366	1.078	1.078	0.319	–0.528	0.183	0.183	0.225	–0.877	–0.166	0.166

words, made up of 300 lines apiece from each of the twelve books—than to *The World's Infancy* (1658), Nicholas Billingsley's 11,111-word versification of the Book of Genesis? Columns A and B show the thirty most common words in descending order of their frequency in the main database. Column C shows their mean frequencies, all represented as percentages of that set, and Column D shows the corresponding standard deviations. Columns F, I, and N show the scores for the whole of Milton's early verse, for our selection from *Paradise Lost*, and for *The World's Infancy*, respectively, and Columns G, J, and O give z-scores representing their divergences from the means of the main set. The z-scores are used to obtain cognate figures for all the words in a hierarchy where the original frequencies fall away sharply from top to bottom.[4] The object is to treat all of these words as markers of potentially equal power in highlighting the differences between one style and another. Columns K and P, respectively, show the differences between the z-scores for Milton and *Paradise Lost*, and those for Milton and *The World's Infancy*.

The next step is to translate the positive and negative measures of difference shown in Columns K and P into absolute differences, as shown in Columns L and Q. By doing so, we obscure some useful stylistic information. But we are now able to derive meaningful totals and means for the whole range of differences. These are shown in L3 and Q3 and in L4 and Q4. A 'delta-score', as I propose to term entries like those in L4 and Q4, can be defined as 'the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text'. (In the current inquiry, the text-groups are authorial. But that could be altered for tasks of non-authorial classification such as the differentiation of genre or era.) The delta-scores of 1.050 and 1.205 show that *Paradise Lost* is less unlike Milton than is *The World's Infancy*. Thirty words, of course, are really too few for our purpose, especially when several of them are pronouns of volatile frequency. But even thirty words are enough to show why the differences we wish to add up and average out must be derived from z-scores and not from the original text-percentages. The text-percentages fall away so rapidly as the list extends downward that even sharp differences among lower-order words would be obliterated, in the total, by those from higher in the order.

Delta-Formel

Although this is a satisfactory outcome, success in a two-horse race does not promise success elsewhere. The addition of further specimens, the complete texts of *Paradise Regained* (15,694 words) and *Samson Agonistes* (12,885 words), each of which behaves as it should, is more encouraging. But the Delta procedure really begins to come into its own when it demonstrates that, although these three of Milton's poems form no part of our Milton-set, they are less different from it than from any other of twenty-five authorial sets. Table 2 shows how the open, multi-author version of the procedure is used to test *Paradise Lost*.

If Columns A–C, where the output is recorded, are passed over for the moment, Table 2 begins like Table 1. Columns D–G show the upper range of the descending hierarchy of common words, standardized

4 An outline of the calculation and use of z-scores can be found in introductory manuals of statistics. But readers in need of such help may be best served by the lucid plain-language account in Kenny (1982, pp. 57–8).

Table 2 First page of 150-word worksheet (*Paradise Lost* as test-piece)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1								**Test-piece**		**Behn**			**Brome**		
2	**MAX**	**1.828**						**150**		**COUNT**		**150**	**COUNT**		**150**
3	**MIN**	**1.023**								**SUM**		**235.2557**	**SUM**		**253.177**
4	**MEAN**	**1.470**								**MEAN**		**1.568**	**MEAN**		**1.688**
5	**STDEV**	**0.176**						**Paradise Lost**		**STDEV**		**1.167**	**STDEV**		**1.467**
6		**OUTPUT**			**DERIVED FROM DATABASE**			**INPUT**							
7		**delta-scores**	**delta z-scores**		**Word**	**Mean**	**SD**	**Score**	**z-score**	**Score**	**z-score**	**Abs. diff.**	**Score**	**z-score**	**Abs. diff.**
8	Behn	1.568	0.560	1	the	4.242	0.630	4.091	–0.239	4.202	–0.064	0.175	3.883	–0.570	0.331
9	Brome	1.688	1.238	2	and	3.770	0.501	4.165	0.789	3.925	0.311	0.478	4.695	1.847	1.058
10	Butler	1.502	0.181	3	of	1.821	0.315	2.769	3.015	1.783	–0.121	3.136	1.229	–1.883	4.898
11	Congreve	1.242	–1.291	4	a	1.601	0.430	0.696	–2.103	1.479	–0.283	1.819	1.750	0.347	2.450
12	Cotton	1.565	0.538	5	to(i)	1.419	0.272	1.289	–0.480	1.331	–0.323	0.157	1.666	0.908	1.387
13	Cowley	1.316	–0.874	6	in(p)	1.358	0.189	1.720	1.916	1.120	–1.264	3.181	1.198	–0.847	2.763
14	Denham	1.344	–0.716	7	his	1.154	0.323	1.532	1.171	0.912	–0.747	1.918	0.978	–0.543	1.713
15	Dorset	1.663	1.098	8	with	1.022	0.208	1.484	2.224	0.944	–0.371	2.595	0.812	–1.006	3.229
16	Dryden	1.393	–0.435	9	to(p)	1.014	0.131	1.245	1.761	0.986	–0.217	1.978	1.026	0.087	1.673
17	Durfey	1.393	–0.436	10	is	0.938	0.312	0.239	–2.238	0.797	–0.452	1.786	1.642	2.254	4.492
18	Gould	1.575	0.596	11	but	0.923	0.195	0.696	–1.167	0.797	–0.648	0.519	1.222	1.536	2.703
19	Marvell	1.367	–0.584	12	he	0.803	0.241	0.703	–0.413	0.792	–0.043	0.370	0.897	0.392	0.805
20	Milton	1.023	–2.536	13	all	0.781	0.193	0.836	0.283	1.179	2.063	1.780	0.840	0.301	0.019
21	Oldham	1.389	–0.456	14	I	0.766	0.391	0.700	–0.171	1.382	1.574	1.745	1.093	0.836	1.007
22	Phillips	1.828	2.033	15	it	0.766	0.239	0.151	–2.575	0.733	–0.138	2.437	1.290	2.196	4.771
23	Prior	1.258	–1.201	16	as	0.710	0.224	0.737	0.119	0.673	–0.167	0.286	0.765	0.246	0.128
24	Radcliffe	1.673	1.153	17	their	0.641	0.237	0.795	0.653	0.355	–1.206	1.859	0.711	0.296	0.357
25	Rochester	1.697	1.292	18	her	0.623	0.336	0.435	–0.560	0.299	–0.961	0.402	0.200	–1.258	0.698
26	Sedley	1.589	0.674	19	not	0.616	0.174	0.847	1.324	0.539	–0.441	1.765	0.989	2.135	0.811
27	Settle	1.412	–0.330	20	be	0.586	0.167	0.401	–1.109	0.617	0.188	1.297	1.016	2.579	3.688
28	Shadwell	1.433	–0.208	21	you	0.580	0.252	0.037	–2.154	1.133	2.196	4.350	0.620	0.157	2.311
29	Swift	1.461	–0.049	22	they	0.564	0.234	0.464	–0.428	0.272	–1.249	0.820	1.049	2.071	2.499
30	Tate	1.330	–0.793	23	for(p)	0.559	0.114	0.000	–4.905	0.475	–0.739	4.167	0.758	1.753	6.658
31	Waller	1.523	0.300	24	by(p)	0.555	0.106	0.689	1.260	0.479	–0.717	1.977	0.569	0.128	1.131
32	Wharton	1.513	0.243	25	my	0.512	0.370	0.258	–0.687	1.221	1.914	2.600	0.339	–0.469	0.218
33				26	we	0.510	0.275	0.265	–0.891	0.290	–0.800	0.091	1.226	2.603	3.494
34	X			27	from	0.500	0.127	0.884	3.019	0.396	–0.816	3.835	0.318	–1.430	4.449
35	Y			28	that(rp)	0.476	0.228	0.313	–0.715	0.636	0.699	1.414	0.968	2.155	2.870
36	Z			29	or	0.471	0.165	0.906	2.636	0.442	–0.175	2.811	0.660	1.145	1.490
37				30	our	0.460	0.268	0.354	–0.397	0.290	–0.634	0.236	0.951	1.835	2.233
38				31	thy	0.451	0.247	0.490	0.158	0.769	1.290	1.132	0.213	–0.961	1.119
39				32	was	0.437	0.140	0.250	–1.333	0.507	0.500	1.833	0.389	–0.340	0.993
40				33	this	0.426	0.095	0.505	0.820	0.304	–1.283	2.103	0.552	1.316	0.496
41				34	when	0.426	0.105	0.284	–1.355	0.544	1.117	2.472	0.345	–0.768	0.587
42				35	are	0.413	0.134	0.063	–2.623	0.382	–0.229	2.394	0.691	2.078	4.701

means for the frequency of each word in our main database, and the corresponding standard deviations. (In its complete form, the table includes the 150 most common words, ranging down to those that occur about once in every thousand in the main database. The words tagged so as to distinguish homographic forms are accompanied by parenthetic abbreviations: (i) for infinitive, (p) for preposition, (rp) for relative pronoun, and (c) for conjunction.) Columns H and I now provide a site for pasting-in the scores for any chosen test-piece and for the z-scores derived by setting those scores against the means and standard deviations given in Columns F and G. In Columns J–L and M–O, respectively, we have the entries for the first two of our twenty-five authorial sets. (In its complete form, the table continues until it includes them all. The vast arithmetical power of spreadsheets such as Microsoft Excel also allows room for other sets to be added as desired.)

Columns J–L (and each of the corresponding trios that follow) give the standardized score for each word in a particular authorial set, the corresponding z-score, and the absolute difference between each z-score and that of the test-piece. Cells L2–L5, O2–O5 (and those corresponding to them in the pages not shown) sum up the columns beneath, giving a count of the number of entries and the sum, mean, and standard deviation of those entries. As in Table 1, these means are our delta-scores. But, refining on Table 1, they are auto-copied across to Column B where each is listed beside the name of the appropriate author. (The entries marked X, Y, and Z at the foot of Column B allow for the addition of other authors or for the testing of control-sets.)

Cell B3 shows the minimum entry in Column B. Glancing down, we see, in Cell B20, that this is Milton's delta-score. On this test, then, *Paradise Lost* differs less from our Milton-set than from any other of our twenty-five authorial sets. With a delta-score of 1.023 on the word-list of 150 (which may be expressed as $\Delta 150 = 1.023$), Milton therefore has the best claim of these twenty-five poets to the authorship of *Paradise Lost*. The strength of the result shown in Cell B20 is reinforced in Cell C20, by far the lowest of a fresh set of 'delta z-scores' derived from the delta-scores in Column B. Milton's 'delta z-score' ($\Delta z150 = -2.536$) diverges so far from the other twenty-four that only one case in a hundred of a normal population could be expected to equal or exceed it. Although it would be unwise to assume that twenty-five sizeable authorial sets constitute a fair sample of seventeenth-century English poetry, there is no obvious reason to insist that they do not. The outcome of our further trials is relevant.

By repeatedly entering new sets of scores in Column H, we can apply the procedure to as many test-pieces as we wish, each test being independent of the rest. The companion paper, 'Questions of authorship', presents a first group of results. These were obtained by applying the Delta procedure (using the scores for all of the 150 most common words) to sixteen long poems by members of our set of twenty-five poets and to another sixteen by poets from beyond that set. Thirty of the thirty-two long poems behaved as they should. Fifteen of the first sixteen attached

themselves firmly to their true authors. Fifteen of the second sixteen showed only weak affinities for authors within the main set, suggesting that their true authors should be sought elsewhere.

The present paper rests upon an extension of the inquiry. It treats of the results obtained by applying the Delta procedure to 200 poems by members of our set of twenty-five poets.[5] The word-lists used fall into a series ranging from the 150 most common down through 120, 100, 80, and 60, to the 40 most common. (The corresponding analysis of further poems by non-members of the set has not yet been undertaken. But the procedure cannot be employed effectively for this further purpose unless the analysis of members of the set is operating at a high level of success. That tends to restrict its usefulness with putative non-members to longer texts where, as will be seen, the success rate for members of the set rises above 90 per cent.)

The choice of the 200 poems was less carefully designed than one might, with hindsight, have desired. The obvious first step was to examine some much shorter poems than those mentioned above and to include enough of them to reduce the risk of sampling errors. The results were favourable enough to encourage a long series of increments and extensions in which more poems by more authors were gradually added. In terms of their length, the 200 poems fall into five bands. A hundred of them are of 500 words or fewer. (Thirty-three of these range from 103 words to 250 and sixty-seven from 251 to 500). The next forty range up from 500 to 1,000 words (twenty of them lying on either side of 750). The remaining sixty fall into three groups of twenty, ranging upward in successive bands to 1,500, 2,000, and beyond. (The last set of twenty long poems includes the sixteen of the companion paper.)

In the matter of authorship, the process of selection was a struggle against the (admittedly fortunate) constraint that poets do not write to our dictates. With some members of the set, no further authentic pieces were to be had. With others there was a dearth of short poems or of long ones. With some there was an unavoidable sameness, with others an extreme diversity. The final group of 200 comprised between twelve and nineteen texts apiece by twelve of our twenty-five poets and nine more by three others. Except for a different selection from *Hudibras*, none of them, in whole or part, figured in the original database against which they were to be tested. All but a dozen were whole poems. To round out the set of twenty in the 1,500 band, there are two separate Cantos from Cotton's *Voyage to Ireland in Burlesque* and two selections from *Paradise Lost*. The latter pair replace the one large selection used at the beginning of this paper.

Probleme mit Texten

The outcome of the whole battery of tests is summarized in Table 3. (The results are also set out, poem by poem, in the Appendix.) In the successive columns making up the left-hand block of the table, the poems are sorted into five bands according to their length. Each cell in this block shows the number of poems whose true authors attained a given rank (out of twenty-five). Thus, in the first row of data, the true author ranked first out of twenty-five for twenty-seven of the 100 poems ranging up to

5 Many of the word-counts derive from texts in the excellent Chadwyck–Healey archive of English poetry, to which my university subscribes. The texts are not used in any other way.

500 words in length; for eighteen of the forty poems ranging from 501 to 1,000 words; for thirteen of the twenty poems ranging from 1,001 to 1,500 words; and so on up to nineteen of the twenty poems of more than 2,000 words. The corresponding columns of the right-hand block express these figures as percentages.

Table 3 Two hundred poems of the late seventeenth century: summary of results of delta-tests

	Poems sorted by length, showing authors' ranks (ex 25)											
	1–500	**501–**	**1001–**	**1501–**	**2001–**	**Totals**	**1–500**	**501–**	**1001–**	**1501–**	**2001–**	**Totals**
	100	**40**	**20**	**20**	**20**	**200**	**%**	**%**	**%**	**%**	**%**	**%**
150 words												
1st	27	18	13	17	19	94	27.0	45.0	65.0	85.0	95.0	47.0
1st–2nd	40	26	14	18	20	118	40.0	65.0	70.0	90.0	100	59.0
1st–5th	67	32	18	20	20	157	67.0	80.0	90.0	100	100	78.5
06 : 10	15	8	2	0	0	25	15.0	20.0	10.0	0	0	12.5
11 : 15	10	0	0	0	0	10	10.0	0	0	0	0	5.0
16 : 20	5	0	0	0	0	5	5.0	0	0	0	0	2.5
21 : 25	3	0	0	0	0	3	3.0	0	0	0	0	1.5
120 words												
1st	31	16	10	15	19	91	31.0	40.0	50.0	75.0	95.0	45.5
1st–5th	62	30	18	20	20	150	62.0	75.0	90.0	100	100	75.0
06 : 10	17	9	2	0	0	28	17.0	22.5	10.0	0	0	14.0
11 : 15	13	1	0	0	0	14	13.0	2.5	0	0	0	7.0
16 : 20	5	0	0	0	0	5	5.0	0	0	0	0	2.5
21 : 25	3	0	0	0	0	3	3.0	0	0	0	0	1.5
100 words												
1st	25	17	8	15	16	81	25.0	42.5	40.0	75.0	80.0	40.5
1st–5th	59	29	18	19	20	145	59.0	72.5	90.0	95.0	100	72.5
06 : 10	18	7	1	1	0	27	18.0	17.5	5.0	5.0	0	13.5
11 : 15	10	4	1	0	0	15	10.0	10.0	5.0	0	0	7.5
16 : 20	11	0	0	0	0	11	11.0	0	0	0	0	5.5
21 : 25	2	0	0	0	0	2	2.0	0	0	0	0	1.0
80 words												
1st	26	12	9	15	16	78	26.0	30.0	45.0	75.0	80.0	39.0
1st–5th	66	29	17	19	20	151	66.0	72.5	85.0	95.0	100	75.5
06 : 10	13	7	2	1	0	23	13.0	17.5	10.0	5.0	0	11.5
11 : 15	14	2	1	0	0	17	14.0	5.0	5.0	0	0	8.5
16 : 20	4	2	0	0	0	6	4.0	5.0	0	0	0	3.0
21 : 25	3	0	0	0	0	3	3.0	0	0	0	0	1.5
60 words												
1st	28	7	5	12	17	69	28.0	17.5	25.0	60.0	85.0	34.5
1st–5th	63	26	17	18	20	144	63.0	65.0	85.0	90.0	100	72.0
06 : 10	22	8	3	2	0	35	22.0	20.0	15.0	10.0	0	17.5
11 : 15	9	5	0	0	0	14	9.0	12.5	0	0	0	7.0
16 : 20	5	0	0	0	0	5	5.0	0	0	0	0	2.5
21 : 25	1	1	0	0	0	2	1.0	2.5	0	0	0	1.0
40 words												
1st	23	10	6	8	17	64	23.0	25.0	30.0	40.0	85.0	32.0
1st–5th	57	29	14	20	19	139	57.0	72.5	70.0	100	95.0	69.5
06 : 10	17	8	4	0	1	30	17.0	20.0	20.0	0	5.0	15.0
11 : 15	13	2	2	0	0	17	13.0	5.0	10.0	0	0	8.5
16 : 20	11	0	0	0	0	11	11.0	0	0	0	0	5.5
21 : 25	2	1	0	0	0	3	2.0	2.5	0	0	0	1.5

Studied from head to foot, the table treats the successive word-lists in descending order from 150 to forty. In each set, the table shows the number of poems for which the true author ranked first, first–fifth, sixth–tenth, and so on. In the top set, the number for which the true author ranked first–second is also given.

The most general results lie in the last column of all. Of the whole 200 poems, a vast miscellany, 47 per cent attach themselves to their true authors when the full word-list of 150 is employed. For 59 per cent of them, the true author ranks either first or second out of twenty-five. For almost 79 per cent, the true author ranks among the first five and for 12.5 per cent among the next five. In only 4 per cent does the true author rank below fifteenth. When the word-list of 120 is employed, the results are a little weaker at almost every point. The lower parts of the table show that continued truncation of the word-list produces a continued deterioration in the results. (Trials in which the word-list was truncated from the top instead of the bottom yielded worse results than any of those shown. The words ranking from 76 to 150, for example, yielded some clean hits but many wild misses.)

If we shift the perspective and weigh up the poems in increasing order of length by moving horizontally across the table, the strongest results of all are for the lists of 150 and 120 on the twenty poems of more than 2,000 words. With nineteen of them, the true author ranks first of twenty-five. With the solitary exception, *The Hind and the Panther*, the true author (John Dryden) ranks second. Although the shorter word-lists yield weaker results than these (save for a minor aberration in the list of 40), the progressive lengthening of the poems always makes for increasing accuracy.

A study of so many specimens justifies some strong conclusions. The overall level of success is impressive because the task of identifying the right candidate from a group of twenty-five offers a much higher level of difficulty than the two- or three-author tasks to which we are accustomed. The unfettered operation of chance would lead, after all, to a roughly equal spread over the several ranks. Two hundred trials would yield only eight cases, not ninety-four, in which the true author ranked first out of twenty-five. In forty cases, not 157, the true author would rank between first and fifth. And in forty cases, not three, the true author would rank between twenty-first and twenty-fifth. Only a genuine authorial factor could yield results like those we have.

It is evident that, with texts of 1,500 words or more, the Delta procedure is effective enough to serve as a direct guide to likely authorship. With problem texts of that length, the procedure has much to offer, especially when an extensive word-list is employed. Even though corroborative evidence would usually be sought, that makes it a worthwhile addition to our scholarly armoury.

Erfolg von Delta

It is also evident that, even with much shorter texts, the Delta procedure is useful in two less direct ways. It makes a basis for selecting a likely group of candidates. The word-lists of 100, 80, and 60 all include the true author among the top five candidates for 85 per cent or more of texts

of above 1,000 words. That supports the use of such Delta trials as a prelude to tests in the 'closed forms' to which I have previously referred. Even more usefully, perhaps, with shorter texts, the Delta procedure helps (as the old song says) to 'ee-lim-inate the negative'. If a putative author does not rank within the top ten of twenty-five candidates, one might demand extremely strong external evidence in his or her favour before discountenancing the doubt so cast. We are thus offered useful negative evidence for cases where it is appropriate. This negative evidence also offers strong general support for the ancient but no longer unchallenged belief that the concept of authorial signatures is well-founded.

How are we to choose the most useful from among these word-lists? Even with the shortest poems, the list of forty yields the least accurate results and should therefore be abandoned. For all groups except the shortest, where some small irregularities appear, the pattern of results improves with each extension of the list. Now, as was noted above, large sets of variables usually yield the most accurate stylistic signatures, possibly as a reflection of the wealth of information they incorporate. But one would scarcely have expected that principle to extend to the point where the list of 150 words still yields some of the best results for poems of as little as 100 words in length.

To move beyond this point, it is necessary to go behind Table 3 and inspect the full authorial hierarchies for the various poems. As the data derived from very short poems become too sparse to be reliable, a statistical artefact intervenes. Of these twenty-five poets, John Milton has the most constrained and strongly delineated stylistic repertoire and, accordingly, moves furthest from the common patterns of the language. Far more often than any of the others, his scores diverge below the norms derived from our main database. (Of the top 100 words, he lies below the norm for seventy-two!) The fact that this makes for many negative z-scores is obliterated when we treat all the scores as absolute. But the fact that many of them are strong divergences produces an unusual pattern of absolute differences. That pattern, as it happens, coincides with one of a very different origin. With very short poems, many of our most common words do not occur at all. Here, again, therefore, the scores often diverge below the norm. Here, again, the divergences yield an unusual pattern—a pattern not unlike Milton's. When the list of forty words is applied to the thirty-three shortest poems, which range in length from 103 to 250 words, Milton's performance is entirely unremarkable. Although he is not the author of any of them, he ranks first for one and between first and fifth for seven. But when the list of 100 words is applied to the same thirty-three texts, he ranks first for seven and between first and fifth for fourteen. With the shortest poem of all, a little song of Congreve's running to only 103 words, the list of forty words puts Rochester at the head of the field. But Milton, who does not rank among the top five candidates, sweeps into first place when the list of 100 is employed. Milton's spurious claim to the authorship of the shortest poems grows even more pervasive when the lists are extended to 120 and 150.

The tests also give Rochester many short poems to whose authorship he has no claim. In this case, however, the progressive extension of the wordlist works differently. With the list of forty words, Rochester ranks first for eight of these thirty-three short poems and between first and fifth for nineteen. With the list of 100 words, he ranks first for six and, once more, between first and fifth for nineteen. The reason this time is not that the scores are unusual. It is rather that they are so consistently normal—so characteristic of the period?—that they are set in low relief. So long as there is a sparsity of information, this pattern of low relief allows many short poems to show a statistical affinity for Rochester. But the longest word-lists, with their richer information, put paid to this false effect. The better delineated frequency-profiles of the longer texts, likewise, are not vulnerable in this way.

Given the weight of misleading influences like these, we must consider how it is possible for the procedure to yield so many accurate results for the thirty-three shortest poems. Even with the list of forty, the true authors rank first for six poems and between first and fifth for seventeen, a score outmatched only by Rochester. With the list of 100, the true authors rank first for eight poems and between first and fifth for twenty-two, outmatching both Rochester and Milton. No other poet of the twenty-five surpasses the constraints of chance.

A poem of only 209 words, *To Alexis, On his saying, I lov'd a Man that talk'd much*, is correctly assigned to Aphra Behn by all six of our word-lists. The poem's 209 word-tokens represent 120 word-types. Of these, only sixty-five word-types and 136 word-tokens lie within the ambit of our 150 most common words. Eighty-five spaces in the hierarchy are not occupied and a further forty spaces each contain a solitary member. Such a list as this does not constitute a frequency-profile in the usual sense of the term. And yet, so the result implies, it best matches the frequency-profile for Aphra Behn's authorial subset in the main database. The absolute z-scores on which the Delta procedure operates have the effect, it seems, of presenting the lower part of this pattern of divergences in almost binary terms. Provided a sufficient number of the occupied spaces in the poem's profile match those where Behn's authorial frequency-profile shows positive divergences and provided a sufficient number of the blank spaces match those where she diverges below the norm, she must (and does) emerge as the top-ranking candidate. This interpretation of the effect complies with the rather hit-or-miss character of the results obtained from the shorter word-lists: a binary hierarchy lacks the subtlety of a true frequency-profile and can easily go amiss.

A more detailed scrutiny of the results reveals some of the limitations of the Delta procedure. The poem-by-poem record shown in the Appendix makes it clear that this is far from a level playing field. The procedure yields a very high success rate with Samuel Butler, whose poems are longer than most, and Katherine Phillips (*Orinda*), whose style is both idiosyncratic and extremely homogeneous. Many of the worst results can be attributed to the fact that, in one way or another, a given poem or group of poems is uncharacteristic of its author. The three love

poems that open Oldham's set, the elegy on Katharine Kingscote, and the epistle to Madam L. E. are all remote from the strong vein of satire for which he is best known and which rightly predominates in his authorial subset. The weak results for Robert Gould's short poems reflect the regrettable fact that his authorial subset is a badly skewed sample of his work. Knowing him only for his long satires, I did not include any of his very diverse short pieces in the original database. The weak results that open Cowley's set are the product of a conscious and (unduly?) successful experiment. The first five of his nineteen poems are the work of his youth, written as much as half a century before the poems for which he is best known. My experience has been that, although authors' styles change over the years, they do not change beyond recognition. This more usual situation is represented here by Waller's *Of the Danger his Majesty . . . Escaped*, which is also a work from the beginning of a long career. But in Cowley's case, the signs of change are so strong as to affect our overall result. For a fine specimen of 'stylochronometry', see Forsyth (1999).

Apart from the very short poems considered earlier, those cases where the procedure gives poor results on poems of more than 500 words represent bad matches—recognizably bad matches—between specimen and authorial subset. They mostly arise from a difficulty encountered by everyone who works in computational stylistics—the fact that authors work at times in very uncharacteristic literary genres. Whereas procedures such as principal component analysis can often overcome the aberrations that arise in this way by absorbing them in the lesser vectors of their output, the single vector of the Delta procedure has no such cushion. With the Delta procedure, an aberrant set of scores is expressed as a lower ranking for the author in question. This being so, it is surprising that the procedure is as resilient as the results demonstrate.

In cases where we are testing the claims of acknowledged candidates, it would usually be possible to foresee this genre-difficulty through our knowledge of their writings. A misleading ranking could be given its due and further tests could be undertaken. But if the Delta procedure is to be of real value in open cases, another answer must be found. Table 4 takes us back to the difference between 'inside' and 'outside' candidates and suggests an answer to this difficulty.

The companion paper, 'Questions of authorship', contrasts the results of the Delta procedure for thirty-two long poems, sixteen by members of our set of twenty-five and sixteen by outsiders. It shows that, when the 150 word-list is employed, the true authors rank first for fifteen of the sixteen poems by insiders and that, in the sixteenth case, the true author ranks second. The sixteen poems by outsiders must, by definition, be 'least unlike' some member of the set of twenty-five authors. They have nowhere else to go. But 'least unlike' need not be 'much like'. The average delta-score for these sixteen is 1.332 as opposed to 1.097 for the sixteen 'insiders'.

Table 4 takes eight poems from each set to reach inside these averages and show how insiders differ from outsiders. (The insiders occupy the

Table 4 Sixteen long poems: summary of results

	Cowley *Davideis* **6812 (sel.)** **delta**			**Waller** *Instructions to Painter* **2606** **delta**			**Dryden** *Absal. and Achit.* **7824** **delta**			**Dryden** *Hind and Panther* **19896** **delta**		
List	**score**	**z-score**	**LU**	**score**	**z-score**	**LU**	**score**	**z-score**	**LU**	**score**	**z-score**	**LU**
150	1.006	2.204	A	1.274	2.209	A	0.993	2.136	A	0.861	1.625	*A
120	1.024	2.486	A	1.393	1.910	A	0.998	2.119	A	0.858	1.798	*A
100	1.030	2.195	A	1.456	1.584	*B	0.965	2.115	A	0.843	1.884	*A2
80	0.934	2.137	A	1.400	1.643	*B	0.966	1.876	A	0.866	1.853	*A
60	0.938	1.932	A	1.459	1.559	A	0.966	1.643	*B	0.815	1.683	*A
	A = Cowley			A = Waller B = Cowley *Waller 2nd			A = Dryden B = Durfey *Dryden 2nd			A = Swift *Dryden 2nd		
	Oldham **Satyr 2** **2210** **delta**			**Gould** **The Presbytery** **4492** **delta**			**Durfey** **The Malecontent** **7817** **delta**			**Swift** **Verses on the Death** **3206** **delta**		
List	**score**	**z-score**	**LU**	**score**	**z-score**	**LU**	**score**	**z-score**	**LU**	**score**	**z-score**	**LU**
150	1.215	3.060	A	1.204	1.915	A	0.745	3.003	A	1.284	2.773	A
120	1.257	3.013	A	1.203	1.911	A	0.746	3.002	A	1.284	2.777	A
100	1.237	2.553	A	1.207	1.568	A	0.733	2.744	A	1.298	2.662	A
80	1.206	2.378	A	1.082	1.704	A	0.741	2.573	A	1.351	2.478	A
60	1.170	2.544	A	1.052	1.418	A	0.668	2.491	A	1.266	2.245	A
	A = Oldham			A = Gould			A = Durfey			A = Swift		
	Fletcher **Purple Island** **5933 (sel.)** **delta**			**Davenant** **Gondibert** **5167 (sel.)** **delta**			**Wild** **Iter Boreale** **3321** **delta**			**Wase** **Divination** **2156** **delta**		
List	**score**	**z-score**	**LU**	**score**	**z-score**	**LU**	**score**	**z-score**	**LU**	**score**	**z-score**	**LU**
150	1.210	1.822	A	1.302	2.185	A	1.324	1.558	A	1.362	1.221	A
120	1.205	1.720	A	1.306	1.786	A	1.314	1.788	B	1.347	1.619	B
100	1.218	1.733	A	1.315	1.505	B	1.297	2.024	C	1.295	1.368	C
80	1.202	1.513	A	1.317	1.508	C	1.246	1.855	C	1.286	1.549	D
60	1.172	1.659	B	1.263	2.033	C	1.142	2.255	A	1.374	1.222	D
	A = Congreve B = Cowley			A = Denham B = Marvell C = Cowley			A = Cowley B = Oldham C = Brome			A = Swift B = Oldham C = Tate D = Prior		
	Pordage **The Medal Revers'd** **3103** **delta**			**Heyrick** **The New Atlantis** **8797 (sel.)** **delta**			**Duke** **Paris to Helena** **3892** **delta**			**Blackmore** **King Arthur** **6986 (sel.)** **delta**		
List	**score**	**z-score**	**LU**	**score**	**z-score**	**LU**	**score**	**z-score**	**LU**	**score**	**z-score**	**LU**
150	1.542	1.799	A	1.162	1.707	A	1.279	1.665	A	1.399	1.526	A
120	1.466	2.114	A	1.183	1.694	B	1.260	1.950	A	1.495	1.517	A
100	1.413	2.491	A	1.210	1.622	B	1.269	1.882	A	1.589	1.487	A
80	1.446	2.503	A	1.230	1.438	C	1.294	1.936	A	1.636	1.312	B
60	1.466	2.201	B	1.119	1.542	C	1.160	2.057	A	1.603	1.330	B
	A = Brome B = Gould			A = Prior B = Durfey C = Tate			A = Behn			A = Milton B = Cowley		

top half of the page.) The table includes the two cases, Dryden's *The Hind and the Panther* and Davenant's *Gondibert*, that give most difficulty. For each poem, the table shows the delta-score on each of the top five word-lists, the corresponding delta z-score from a set of twenty-five, and in the column headed 'LU' (for 'least unlike'), a set of codes for the top-ranking candidates. Cowley's *Davideis*, the first entry, shows delta-scores approximating to 1.000, delta z-scores ranging down to almost –2.5, and a consistent top-ranking for Cowley.

A close study of Table 4 shows several contrasting tendencies, none of which is absolute, in the two sets of eight poems. The delta-scores for the insiders run lower than those for the outsiders. The delta z-scores run much more strongly into the negative. (Their special value is to offset the fact that delta-scores rise rapidly for shorter texts. The delta-scores shown here, averaging 1.075 for insiders and 1.318 for outsiders, are consistent with the length of these poems.) For the insiders, the lowest delta-scores and, accordingly, the most strongly negative of the delta z-scores tend to emerge from the longer word-lists, where the information used is richer. And the insiders tend to sit consistently with a single candidate. When all of these tendencies are weighed up, none of the poems by 'outsiders' can easily be taken for the work of any member of the set of twenty-five poets. In the other set, the overall case for each of the true authors is impressive for all but *The Hind and the Panther.* The particular reasons why that poem breaks the pattern are examined in 'Questions of authorship' and a second round of testing rectifies the first.

'Questions of authorship' includes a bar-graph in which the delta z-scores for my original set of thirty-two long poems on the 150 word-list shows that a threshold of –1.9 neatly separates the two sub-sets of sixteen poems save for the two exceptions I have mentioned. (In a normal population, a z-score of –1.9 separates around 3 per cent of cases from the remainder. Two exceptions out of thirty-two are by no means unexpected.) The possibility of using just such a bar-graph as a grid for testing specimens of unknown authorship was not to be ignored.

But the addition of only four more long poems by 'insiders' to round out the new group of twenty poems of above 2,000 words yielded another exception. John Oldham's pompous 2,237-word eulogy *Upon the Works of Ben. Johnson* does register as his on the longer word-lists and, after a fleeting affinity with Dryden, returns to him on the shorter word-lists. But the delta z-score for the list of 150 is only –1.679, far below the threshold of –1.9. A poem that opens with the invocation 'Great Thou! Whom it is a crime almost to dare to praise' is likely to differ in many ways from the characteristic work of this harsh and argumentative satirist. Although it is reassuring to find that, even here, the Delta procedure does identify the true author, the delta z-score is a sharp reminder that the system for distinguishing between insiders and outsiders is not foolproof. It behoves us, as always, to remember that, by relying on statistical analysis, even in this simple form, we are dealing in probabilities and not in absolutes.

With this necessary proviso and without forgetting the other limitations we have observed, it is clear that the Delta procedure satisfies the

purposes enunciated at the beginning of this paper. Even with very short texts, it is more successful than might have been expected at the 'open game' of picking out the most likely set of candidates from a large group. By comparison with the results described (my own among them) in the extensive trials of Forsyth and Holmes (1996), it is extremely accurate in singling out the true author of texts of more than 1,500 words in length. (That claim is strengthened by the fact that it is much more difficult to identify the true author in a field of twenty-five candidates than in a comparison of two candidates or three.) And it shows promise as a means of indicating that the true author of a given text may lie beyond a current set of candidates, a task we have not hitherto accomplished. How is it that such a primitive statistical instrument can satisfy these purposes? The answer must lie, I believe, in areas where we are still extremely ignorant—in the communicative resilience of the language and the astonishing force of human individuality.

Delta Effektivität

Acknowledgements

The research on which this paper is based has been generously supported by the Australian Research Council and the University of Newcastle. I am also indebted to Harold Love of Monash University and to my colleagues in the Centre for Literary and Linguistic Computing at Newcastle.

References

Bailey, R. W. (1979). Authorship attribution in a forensic setting. In Ager, D. E., Knowles, F. E., and Smith, J. (eds), *Advances in Computer-aided Literary and Linguistic Research: Proceedings of the Fifth International Symposium on Computers in Literary and Linguistic Research*. Birmingham: John Goodman, pp. 1–15.

Binongo, J. N. G. and Smith, M. W. A. (1999). The application of principal component analysis to stylometry. *Literary and Linguistic Computing*, **14:** 445–65.

Burrows, J. (2003). Questions of authorship: attribution and beyond. *Computers and the Humanities*, **37:** 1–26.

Burrows, J. and Craig, H. (2001). Lucy Hutchinson and the authorship of two seventeenth-century poems: a computational approach. *The Seventeenth Century*, **16:** 259–82.

Forsyth, R. S. (1999). Stylochronometry with substrings, or: a poet young and old. *Literary and Linguistic Computing*, **14:** 467–77.

Forsyth, R. S. and Holmes, D. I. (1996). Feature-finding for text classification. *Literary and Linguistic Computing*, **11:** 163–74.

Holmes, D. (2001). A widow and her soldier: stylometry and the American Civil War. *Literary and Linguistic Computing*, **16:** 403–20.

Kenny, A. (1982). *The Computation of Style: an Introduction to Statistics for Students of Literature and Humanities*. Oxford: Pergamon.

Mosteller, F. and Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist Papers* (2nd edn, 1984). New York: Springer.

Tweedie, F., Holmes, D., and Corns, T. (1998). The provenance of *De Doctrina Christiana*, attributed to John Milton: a statistical investigation. *Literary and Linguistic Computing*, **13:** 77–87.

Appendix

Table A1 Two hundred long poems of the late seventeenth century: list of texts with delta-ranks

		Word-list						
		150	120	100	80	60	40	
	Length	Author's rank (ex 25)						
Behn								
a	248	1	1	3	3	3	3	*A Pindaric to Mr. P. who sings finely*
b	500	7	3	7	5	5	5	*On the Authour of that Excellent Book . . . The Way to Health* [etc.]
c	559	7	5	10	10	11	8	*To the Honorable Sir Francis Fane, on his Excellent Play, The Sacrifice*
d	263	12	15	14	15	12	17	*A Satyr on Doctor Dryden*
e	293	12	10	20	22	19	24	*To Alexis in Answer to his Poem against Fruition. Ode*
f	209	2	1	1	1	1	1	*To Alexis, On his saying, I lov'd a Man that talk'd much*
g	273	16	11	9	11	9	15	*To Amintas, Upon reading the Lives of some of the Romans*
h	416	6	3	3	6	4	3	*On the first discovery of falseness in Amintas*
I	454	3	2	3	3	4	5	*On the death of E. Waller, Esq;*
j	161	3	3	3	2	4	4	*Verses design'd by Mrs. A. Behn, to be sent to a fair lady . . . Left unfinish'd*
k	199	5	2	1	1	1	5	*On a Pin that hurt Aminta's Eye*
l	282	11	3	4	4	7	3	*A Letter to the Earl of Kildare, disswading him from marrying Moll Howard*
m	818	1	2	1	2	2	2	*On Desire. A Pindarick*
n	761	7	6	6	3	2	1	*A Pindaric Poem to the Reverend Doctor Burnet*
o	16419	1	1	2	2	3	2	*A Voyage to the Island of Love*
Brome								
a	379	1	1	1	1	1	1	*The Answer (Stay, stay, prate no more)*
b	437	1	1	1	1	1	1	*The Leveller*
c	366	1	1	1	1	1	1	*The Polititian*
d	428	1	1	1	1	1	1	*On Sir G. B., his defeat*
e	453	3	4	4	3	4	4	*On a Butcher's Dog*
f	325	1	2	2	1	1	1	*Palinode*
g	436	1	3	9	8	7	10	*To a Potting Priest upon a quarrel*
h	364	5	6	7	5	5	4	*To the Meritoriously Honorable Lord Chiefe Justice of the Kings bench*
I	479	7	10	16	8	1	13	*Upon the miscarrier of Letters betwixt his Friend and him*
j	376	2	2	3	1	2	1	*To his friend Mr. I. W. on his translation of a romance*
k	241	1	1	1	1	2	7	*Upon the Kings imprisonment*
l	693	1	1	1	2	2	5	*The Answer (Did I not know thee friend)*
m	645	4	9	11	6	6	9	*Upon the Death of that Reverend and learned Divine, Mr. Josias Shute*
n	761	4	4	7	11	14	1	*The Satyr of Money*
o	767	2	3	3	3	4	9	*To C. S. Esquire (Dear Charles, I am thus far come)*
p	1385	1	2	2	2	2	9	*The Answer (My Friend, in troth, I am glad to hear)*
Butler								
a	1217	1	1	1	1	2	1	*Satyr upon Plagiaries*
b	1401	1	1	1	3	2	2	*Satyr upon the Weakness and Misery of Man*
c	1528	1	1	1	1	2	3	*Satyr upon the Licentious Age of King Charles the 2d*
d	1720	1	1	1	1	2	1	*Upon a Hypocritical Nonconformist*
e	1950	1	1	1	1	1	1	*Satyr upon the Imperfections and Abuse of Human Learning*, Pts i and ii, 1–72
f	1621	1	1	1	1	2	1	*Hudibras*, the third part, Canto ii, 1–266
g	615	1	1	1	1	1	1	*Satyr upon Gaming*
h	604	1	1	1	1	4	6	*A Panegyric upon Sir John Denham's Recovery from his Madness*
I	705	1	1	3	4	6	3	*Satyr upon Drunkenness*
j	962	1	1	1	1	1	2	*Satyr upon Marriage*
k	813	1	1	1	1	1	1	*Upon Philip Nye's Thanksgiving Beard*
l	844	1	1	1	1	2	1	*Upon Modern Critics*

Table A1 (*continued*)

		Word-list						
		150	120	100	80	60	40	
	Length	Author's rank (ex 25)						
Congreve								
a	203	2	2	4	2	2	2	*Paraphrase upon Horace*. Ode xix., Lib. 1
b	247	1	1	1	1	2	5	*The Reconciliation*
c	313	17	18	16	4	10	12	*To Mr. Dryden on his Translation of Persius*
d	260	11	12	9	12	6	15	*Prologue to Pyrrhus, King of Epirus*
e	400	2	2	5	3	3	1	*Doris*
f	935	2	2	4	5	11	14	*Of Pleasing. An Epistle to Sir Richard Temple*
g	133	2	2	2	3	6	8	*Song (As Amoret and Thyrsis, lay)*
h	337	3	4	8	9	7	6	*Buxom Joan of Lymas's Love for a Jolly Sailer*
I	103	3	3	3	4	4	5	*The Decay, a Song*
j	321	2	2	3	3	4	2	*A Poem in Praise of the Author*
k	403	5	7	11	10	10	16	*Prologue, written for Mr. Haynes*
l	617	3	3	2	3	2	2	*A Letter from Mr. Congreve to the Lord Viscount Cobham*
m	1349	2	2	2	1	1	2	*The Tears of Amaryllis and Amyntas, a Pastoral*
n	1376	1	1	1	1	2	2	*An Impossible Thing*
Cotton								
a	174	5	11	7	5	10	13	*Elegy*
b	212	16	20	19	11	14	3	*Advice*
c	168	1	1	2	1	1	2	*Epigramme de Monsieur des-Portes*
d	219	5	11	4	5	6	13	*Ode de Monsieur Racan*
e	292	1	1	1	2	1	4	*To my dear and most worthy friend, Mr. Izaac Walton*
f	351	3	3	1	2	2	2	*The World. Ode*
g	345	1	1	1	1	1	1	*A Journey into the Peak*
h	1959	1	1	1	1	1	3	*Burlesque. Upon the Great Frost*
I	1830	2	3	3	4	5	5	*A Voyage to Ireland in Burlesque*. Canto 2
j	1543	1	1	1	1	1	1	*A Voyage to Ireland in Burlesque*. Canto 3
k	1537	1	1	1	1	1	3	*Epistle to Sir Clifford Clifton*
l	789	3	7	7	6	8	3	*On Tobacco*
m	816	1	3	4	3	4	4	*Eclogue*
n	664	4	5	3	6	7	7	*On Marriott*
Cowley								
a	393	4	6	6	4	3	2	*On his Majesties returne out of Scotland*
b	443	9	12	16	12	10	16	*A Poeticall Revenge*
c	256	21	22	20	22	15	18	*To his very much honour'd Godfather, Master A. B.*
d	294	23	21	19	20	12	20	*An Elegie on the death of Mris Anne Whitfield*
e	217	22	17	17	19	19	20	*An Answer to an Invitation to Cambridge*
f	469	2	1	1	1	1	2	*A Discourse By way of Vision Concerning the Government of O. Cromwell*
g	432	3	10	12	3	4	4	*It is a Truth so certain and so clear*
h	312	2	1	2	1	1	3	*When, Lo, e're the last words were fully spoke*
I	482	1	1	1	1	1	1	*A paraphrase on an Ode in Horace's third Book*
j	323	10	9	12	12	9	12	*The Motto*
k	328	1	1	3	3	11	9	*Life and Fame*
l	357	8	14	16	12	13	17	*Life*
m	933	1	1	1	1	1	1	*The First Nemaean Ode of Pindar*
n	772	2	1	1	1	2	2	*Hymn. To Light*
o	1291	1	1	1	1	2	2	*The Second Olympique Ode of Pindar*
p	1182	9	10	9	11	7	14	*The Complaint*
q	1315	1	1	2	2	1	1	*To the Royal Society*
r	1897	1	1	1	1	1	1	*The Plagues of Egypt*, ll, 1–262
s	6812	1	1	1	1	1	1	*Davideis*, Book ii

Table A1 (*continued*)

		Word-list						
		150	**120**	**100**	**80**	**60**	**40**	
	Length	**Author's rank (ex 25)**						
Gould								
a	221	18	20	21	20	20	20	*The Dream*
b	219	4	8	6	6	6	9	*To Mr. Giles Frost*
c	216	14	10	19	18	18	10	*To an unknown Relation; Hearing that he was happily Married*
d	219	15	16	16	12	13	18	*To Mr. Trowe, on the Death of Madam Goddard*
e	176	19	22	21	22	23	18	*To Sir James Long, Baronet, with [an] Elegy on his Brother*
f	184	3	6	10	5	1	6	*On Good Friday*
g	612	8	10	13	16	21	22	*Silvia in the Country*
h	616	1	1	1	2	5	7	*To Julian, Secretary to the Muses, on his Confinement*
I	651	2	2	2	6	11	12	*To Madam B., occasion'd by a Copy of Verses*
j	904	2	2	3	1	5	8	*To Mr. Lowin, from the Country*
k	786	2	6	5	3	6	4	*Of Adorissa's Second Marriage with Mr. Grevil*
l	908	7	12	15	16	10	10	*To the Memory of ... Colonel Edward Cooke*
m	6020	1	1	1	1	1	1	*To the Society of the Beaux Esprits. Ode*
n	4019	1	1	1	1	1	1	*The Play-House, a Satyr. The Second Part*
o	1988	1	3	3	3	6	5	*The Step-Mother*, ll, 1–245
p	1995	1	4	9	10	10	5	*A Glance at Fanaticism*, ll, 1–242
q	4057	1	1	1	1	1	1	*A Satyr against Man. The First Part*
r	4492	1	1	1	1	1	1	*Presbytery Rough-Drawn. A Satyr*
Oldham								
a	568	7	8	10	9	7	5	*Upon a Lady*
b	577	9	7	7	8	5	5	*Ovid's Elegies imitated*: Book ii, Elegy 4
c	652	8	9	11	13	13	5	*The Dream*
d	887	1	2	1	1	2	3	*Upon the Marriage of the Prince of Orange with the Lady Mary*
e	266	13	14	15	14	14	11	*The Parting*
f	409	10	10	10	10	14	4	*On the Death of Mrs. Katharine Kingscote, a Child of Excellent Parts*
g	1298	1	2	4	2	2	1	*Satyr upon a Woman*
h	1710	1	1	1	1	3	3	*An Imitation of Horace*, Book I, Satyr 9
I	1943	1	1	1	1	1	2	*A Letter from the Country to a Friend in Town*
j	1726	1	1	1	2	1	2	*A Satyr address'd to a Friend that is about to leave the University*
k	1254	4	8	14	10	9	11	*To Madam L. E., upon her Recovery from a late Sickness*
l	3795	1	1	1	1	1	1	A *Satyr in imitation of the Third of Juvenal*
m	2237	1	1	4	3	1	1	*Upon the Works of Ben. Johnson*
n	1503	1	1	1	1	2	3	*Counterpart to the Satyr against Vertue*
o	2210	1	1	1	1	1	1	*Satyrs upon the Jesuits*. Satyr ii
p	3378	1	1	1	1	1	1	*The Eighth Satyr of Monsieur Boileau*
Phillips								
a	229	9	1	2	1	2	1	*To her Royal Highness, the Dutchesse of Yorke*
b	242	6	8	8	7	6	1	*To the Right Honoble Alice, Countess of Carberry*
c	181	1	1	2	1	1	1	*Friendship's Mysterys*
d	297	4	11	14	10	10	17	*To Mr. Henry Vaughan, Silurist, on his Poems*
e	238	1	1	1	1	1	3	*To Antenor, on a paper of mine*
f	344	1	1	1	1	1	1	*Rosania's private Marriage*
g	358	1	1	1	2	3	8	*To Mrs. Mary Awbrey at parting*
h	415	2	6	7	7	8	1	*Friendship*
I	190	2	1	2	2	1	1	*Rosania to Lucasia on her Letters*
j	445	4	3	3	2	1	5	*On the Death of Sir Walter Lloid*
k	299	4	1	1	2	1	1	*To the Right Honourable the Countess of Cork*
l	356	1	1	1	1	1	1	*To the Lady Mary Butler, at her marriage*
m	667	1	1	1	1	1	2	*L'accord du bien*
n	554	1	2	1	1	1	2	*The Soule*
o	563	7	9	9	5	9	1	*An Ode upon Retirement*
p	1210	1	1	1	1	1	1	*To the Rt. Hon. the Lady E. C.*

Table A1 (*continued*)

		Word-list						
		150	120	100	80	60	40	
	Length	Author's rank (ex 25)						
Prior								
a	303	4	2	1	1	1	2	*The Lady's Looking-Glass*
b	336	2	2	2	1	1	2	*Prologue, spoken at the Court before the Queen*
c	541	4	5	3	3	2	3	*An English Padlock*
d	364	15	11	9	12	9	9	*Epilogue to Phaedra*
e	192	4	2	2	2	2	1	*Seeing the Duke of Ormond's Picture at Sir Godfrey Kneller's*
f	290	3	3	4	3	2	3	*The Chameleon*
g	321	3	3	2	2	2	2	*Love disarm'd*
h	156	3	2	2	2	2	2	*Cloe hunting*
I	203	3	1	1	1	1	6	*Horace Lib. 1 Ep. ix imitated (To Mr. Harley)*
j	245	5	4	3	3	6	6	*To Mr. Harley, wounded by Guiscard*
k	242	7	5	4	3	4	12	*True Statesmen*
l	149	15	13	12	11	9	11	*A Lover's Anger*
m	933	2	3	3	2	1	1	*Celia to Damon*
n	1112	4	3	3	2	2	5	*The Ladle*
o	1115	7	5	5	4	3	10	*Paolo Purganti*
p	1615	5	3	2	1	1	1	*Alma: First Canto*
q	2697	1	1	1	1	1	1	*An Ode humbly inscrib'd to the Queen*
r	6033	1	1	1	1	1	1	*Henry and Emma*
Swift								
a	310	1	1	6	4	4	7	*The Discovery*
b	316	2	2	3	4	3	2	*The History of Vanbrug's House*
c	406	7	7	6	4	4	6	*Apollo Outwitted*
d	489	1	1	1	2	3	3	*To Lord Harley . . . on his Marriage*
e	354	2	2	4	2	3	3	*The Faggot*
f	535	2	1	1	2	2	1	*The Storm: Minerva's Petition*
g	408	6	5	6	7	3	15	*The First of April: to Mrs. E. C.*
h	366	8	9	11	12	9	11	*A Receipt to Restore Stella's Youth*
I	894	1	1	1	1	2	2	*To Stella, who Collected and Transcrib'd his Poems*
j	823	1	1	1	2	2	2	*An Epistle upon an Epistle*
k	899	1	1	2	3	4	4	*The Lady's Dressing Room*
l	1023	3	2	2	6	4	3	*Ode to the King on his Irish Expedition*
m	1210	5	4	4	4	8	6	*Occasion'd by Sir William Temple's Late Illness and Recovery*
n	1243	1	1	1	1	1	3	*A Libel on the Reverend Dr. Delany and . . . John, Lord Carteret*
o	1093	1	1	1	1	1	1	*To Dr. Delany on the Libels Writ against Him*
p	1433	1	1	2	2	4	4	*The Beasts' Confession*
q	3381	1	1	1	1	1	1	*On Poetry: A Rhapsody*
r	3206	1	1	1	1	1	1	*Verses on the Death of Dr. Swift D.S.P.D.*
Waller								
a	321	3	3	4	3	4	1	*To my Lord of Falkland*
b	554	1	1	1	2	2	1	*Of the Queen*
c	388	1	1	1	1	1	1	*To my Lady Morton, on New-Year's Day, 1650*
d	327	4	6	5	11	19	21	*To Sir William Davenant upon his two first Books of Gondibert*
e	247	1	1	1	3	2	6	*To my ... friend Master Evelyn upon his Translation of Lucretius*
f	263	1	1	2	2	1	1	*Upon the late Storm and the Death of his Highness*
g	202	1	2	1	1	2	2	*Of English Verse*
h	276	6	9	10	8	9	8	*Epitaph on Colonel Charles Cavendish*
I	330	14	11	14	7	6	14	*To the Prince of Orange, 1677*
j	440	1	1	1	1	1	1	*Upon the Earl of Roscommon's Translation of Horace*
k	336	6	13	15	12	4	10	*To Mr. Creech on his Translation of Lucretius*
l	422	1	1	2	2	3	2	*A Presage of the Ruin of the Turkish Empire*

Table A1 (*continued*)

		Word-list						
		150	**120**	**100**	**80**	**60**	**40**	
	Length	**Author's rank (ex 25)**						
m	1777	3	3	5	4	1	3	*The Battle of the Summer Islands*
n	1322	1	3	2	1	5	9	*Of the Danger his Majesty (being Prince) Escaped*
o	2606	1	1	2	2	1	6	*Instructions to a Painter*
Others								
a	1170	1	1	1	1	2	1	Dryden: *Heroic Stanzas consecrated to the Memory of Oliver*
b	1650	1	1	1	1	1	1	Dryden: *Epistle the Fifteenth*
c	7824	1	1	1	1	2	1	Dryden: *Absalom and Achitophel*
d	19896	2	2	2	2	2	2	Dryden: *The Hind and the Panther*
e	1926	1	1	1	1	1	1	Milton: *Paradise Lost*, Book 1, ll, 201–457
f	1867	1	1	1	1	1	3	Milton: *Paradise Lost*, Book 3, ll, 201–447
g	15694	1	1	1	1	1	1	Milton: *Paradise Regained*
h	12885	1	1	1	1	1	1	Milton: *Samson Agonistes*
I	7817	1	1	1	1	1	1	Durfey: *The Malecontent*