All the Way Through: Testing for Authorship in Different Frequency Strata

John Burrows
University of Newcastle, Australia

Abstract

This article describes the operation of two new tests of authorship and offers some results. Both tests rely on controlled contrasts of word-frequency and both exclude the very common words, which have been put to such good use in recent years. One test treats of words used with some consistency by a target-author but more sporadically by others. The second treats of words used sporadically by the target-author but not by most others. (The inclusion of words that some other authors use avoids the strict constraint that has impoverished this form of evidence.) In suitable cases, both tests prove very accurate. The fact that evidence of authorship can be detected in these three distinct frequency-strata helps to explain why such tests should work at all and so encourages the development of even better ones.

Correspondence:
John Burrows,
Centre for Literary and Linguistic Computing,
University of Newcastle,
Callaghan, NSW 2308,
Australia.
E-mail:
john.burrows@netcentral.com.au

1 Rationale

Computational stylistics deals in patterns derived from the relative frequencies of many words across a chosen range of texts. In recent years, much attention has been given, in this area of scholarly inquiry, to the words that occur most often. These, of course, are words that every writer of English will call upon in almost any form of written expression. We now know that there are strong concomitant variations of frequency among many of these words. The consistency of the variations is usually sufficient to bear the weight of statistical analysis. This makes it possible, among other things, to distinguish the writings of different authors from each other and, accordingly, to test the authorship of doubtful texts.

Such idiosyncratic patterns of occurrence in the top stratum of frequency are scarcely apparent to a reader. A given writer's habitual choices among comparatively rare words, on the other hand, are easily seen, often attract comment, and lend themselves to parody. In the large area between the extremes of ubiquity and rarity, it is also possible to identify many words (not all of them unusual) that a given writer employs with some consistency while most others do not. Now if, in comparable bodies of work, some writers use certain words that others do not—and the more so if those words recur—it seems fair to suppose that the users are the most likely to employ them in new specimens of their work. It would follow that each writer's use of such words might yield a distinctive profile.

If the words that make up the second and third of these frequency strata are to yield evidence of authorship, appropriate statistical procedures must be brought to bear. In the third stratum, the inherent fragility of low-frequency statistics is made worse, in our field of inquiry, by a tendency to focus on words not used by anyone except the target-author. The presence of such words in a given

Literary and Linguistic Computing Vol. 22, No. 1, 2007.
doi:10.1093/llc/fqi067 Advance Access Published on 6 January 2006

text can undoubtedly be of evidential value. Their absence may easily be adventitious. And they are usually too few to bear much weight statistically. It is nevertheless possible, as I hope to show, to support (not displace) this form of evidence by relaxing one's stipulations and adding words that few (rather than no) others use. The evidential power of each word is thus reduced but there is more room for a cumulative effect.

In studies of putative authorship, the middle frequency stratum has scarcely been addressed. A body of words, most of them lexical, that we take up and discard according to the needs of a topic or an occasion does not seem likely to offer useful evidence of authorship. Among these words, however, there are many that respond to simple rules of consistency and contrast. Once the words that almost everybody uses are excluded, it is possible to identify words that many writers seldom use but that recur, with some consistency, in a wide-ranging set of work by a given author. Not all of them will reappear in everything that he or she writes; but it seems that more of them can be expected here than in the work of other writers.

Statistical tests directed at each of the three strata are proving extremely accurate in identifying the true author of authentic texts 2,000 words or more in length. The level of accuracy gradually falls away with shorter texts. The errors that do arise, even with longer texts, differ from stratum to stratum. The evidence of the most frequent words can be distorted when an author makes an uncharacteristic choice of genre or literary form, as when Henry Fielding turns his hand to verse. The evidence of the middle range of words can be distorted by a radically different choice of subject, as when Aphra Behn turns her mind from love to death. The evidence of the unusual words is susceptible to aberration as when circumstance carries a particular word into a little flurry of occurrences. Some judicious culling of the word-list alleviates these difficulties. But, while the tests are still being developed, it is better not to interfere with the evidence.

Since there need be no overlapping among the words tested in each stratum, it might appear that the three sorts of test are independent of each other. While that is an important part-truth, it is offset, to an extent we do not much understand, by innumerable links, overt and covert, among the occurrence-rates of different words. An emphasis on the feminine, for example, is to be observed at *every* level of the vocabulary of most female writers. It is present again, in an altered guise, in the vocabulary of Rochester's literary circle. Despite this restriction of their apparent independence, a concurrence in the test-results from the different strata clearly strengthens a given outcome.

Why should such tests work as accurately and reliably as they do? Provided they are properly used, the proper answer, I believe, is 'How could it be otherwise?' Any writer's vocabulary is a selection from the full resources of a given language as used at that time. His or her preferences will reflect such differentiae as level of education, gender, chosen audience, topics of customary choice, and so on. If it is an international language like English, there will be signs of a given national variety. The writer's preferences will also reflect idiosyncrasies too subtle for such broad categories. Such a set of preferences will amount, in short, to one major facet of a Saussurian *parole*, drawn from the larger resources of the *langue* itself. As such, like every other meaningful aspect of our behaviour, they will display not only an underlying likeness, greater or lesser, to our various fellows but also our differences from them. Whether that line of thought is an incipient theory or merely an idea of a certain generality, I am too simple an empiricist to judge. I note, however, that de Saussure's distinction is acceptable to Chomsky, whose distinction between 'competence' and 'performance' is germane. That, I take it, should put it in good standing among adherents of the high-priori sort of language study. Yet, even if de Saussure were denied that privilege, one might still do worse than follow him and bring a little further evidence to show that (in written as well as spoken language) his distinction is as useful as it is plausible.

Let me put it boldly in the form of a postulate. Evidence of authorship pervades whatever anybody writes. Provided appropriate procedures are employed in the analysis of an appropriate set of texts, it can almost always be elicited. It is inherent,

however, not merely in statistical principle but in human behaviour at large, that such evidence cannot be absolute. The consistencies we observe are trends, not universals. Our many stabilities are offset by our capacity for change.

2 Texts to be Assessed

Some of the procedures to be described here are new and need extensive testing. For the more limited purposes of this introductory sketch, the putative authorship of eight poems of the English Restoration era will be assessed. Three are of unquestioned authorship. These are Edmund Waller's 'On the Danger his Majesty (being Prince) Escaped' and 'Instructions to a Painter' and Andrew Marvell's 'The First Anniversary of the Government under O. C.' The other five poems are rejoinders to Waller's 'Instructions,' political satires treating harshly of James, Duke of York (Waller's hero of the moment) and the conduct of naval affairs. Presumably because an author of such work might well be tried for treason, all five were published anonymously and their authorship is still not definitely known. But the 'Last Instructions to a Painter' is accepted as Marvell's. Current scholarly opinion, as reflected in the most recent edition of Marvell's poetry (Smith, 2003) favours him as author of the 'Second' and the 'Third Advice to a Painter.' (See also Patterson, 2000.) The authorship of the 'Fourth' and 'Fifth' remains open but the idea that they may be Marvell's has no support.[1] The case is set out more fully in a recent article (Burrows, 2005), where the evidence of the very common words is brought to bear. The series of tests undertaken below begins with a specimen of the results offered by the very common words.

These eight poems have particular advantages for assessing the effectiveness of tests directed at the less common words. As to Waller, the experiment is fairly straightforward. Two of the eight poems are his, one is not, and the other five cannot be. All eight treat subjects that he favours and (though some are hostile) in a manner broadly related to his. As to Marvell, the matter is more difficult because these political poems stand well apart from most of his acknowledged work. Yet the question is worth asking. Can the tests we shall be examining offer results that accord with scholarly opinion? That would entail an outcome in which the poem known to be Marvell's and the three considered his were distinguished from the two by Waller and the two by unidentified poets. When the proposed contest between Waller and Marvell is resolved, we shall consider whether the authorship of these poems can also be established in a contest where many other poets of the period are introduced.

For some of the necessary comparisons, I have had recourse to a large and diverse database of half a million words of Restoration poetry.[2] A further forty-one independent texts have also been introduced. Twenty-one are poems (or long excerpts from poems) by authors included in the main database. The remaining twenty are by Restoration poets who are not members of the main set.[3]

All the texts have been modernized to overcome the vagaries of seventeenth-century spelling. Contractions like 'I'll' and 'don't' have been expanded. A few words are tagged so as to distinguish among their main grammatical functions.

3 The Evidence of very Common Words: The Delta Test

In the multivariate statistical procedures used to elicit intelligible patterns in the frequency-distribution of the very common words, the texts are specimens and the words are the variables under scrutiny. To allow for differences in length between text and text, the raw frequencies for each word-type are standardized, usually as percentages of all the word-tokens in each text.[4] The standardized frequencies are arranged in descending order as a frequency-profile for each text in turn.

The Delta procedure compares the upper range of the frequency-profile of a given text with those of many authors and shows which of them is least unlike it. The operation of the procedure is described elsewhere and a large body of results is shown. (Burrows, 2002, 2003; Hoover, 2004)[5]. Especially with texts of more than 2,000 words, the results point towards the more likely candidates and allow the elimination of the more unlikely. The errors that occur from time to time stem from

unusually strong changes in an author's style, as between the early and the late work of Cowley; as between the satires and the lyrics of Robert Gould; or as when a gifted translator like Dryden submerges many of his usual stylistic propensities as he tries to catch the spirit of his foreign model.

The outcome of each Delta test stands free of all the others in the sense that each specimen text, in turn, is matched against the same collection of authorial sets. If those sets are changed, the various specimens are all affected; but, in contrast to other procedures like cluster analysis and principal component analysis, the specimens tested do not affect each other.

A 'delta-score', like those shown under that heading in Table 1, 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'. When they are calculated, as here, for a sufficient number of authors, they can be ranked and 'delta z-scores' can be derived to allow more meaningful comparisons among the results for different texts.[6]

Table 1 shows that Waller and Marvell, respectively, rank first of the twenty-five authors for each of the three unquestioned poems. Marvell also ranks first for the 'Last Instructions' and for the 'Second' and the 'Third Advice'. Almost all the scores are strong enough to bear real weight and at no point is there a close contest between Waller and Marvell. Marvell, so it proves, ranks well down the list of twenty-five for the two 'Advices' thought not to be his work. Robert Gould, much too young to be a contender, ranks first for the 'Fourth Advice', followed by Denham, whose claim can be taken seriously. As for the 'Fifth Advice', a close contest among several unlikely candidates is a typical result in cases where the true author lies outside our group of twenty-five. The weak delta z-scores at the head of the list also support that possibility. Table 1, in short, is a fair specimen of the evidence in the article mentioned (Burrows, 2005) and the results are entirely in keeping with current literary scholarship. Taking all this as our basis, we can go on to consider two new approaches to such matters.

4 The Evidence of Less Common Words: The Zeta and Iota Tests

4.1 One on one: Waller versus Marvell; Marvell versus Waller

The Delta procedure focuses on a text and seeks to identify the correct one of many possible authors. The new 'Zeta' and 'Iota' tests, on which I have been working in recent months, focus on a single author and seek to identify which of many texts are most likely to be his or hers. The point of departure for this series of tests is the complete word-list for a large sample of a particular author's work. The 13,838 word-tokens of the Waller set used here embrace 2,876 distinct word-types. (The 20,151 of Marvell embrace 4,323.) Corresponding lists are then established and tabulated, showing the incidence of these same word-types in each of the other members of a large group of authorial sets and also in such independent texts as are to be assessed. In my experiments to date, around 10,000 word-tokens seem to suffice as a reliable minimum for an authorial set, 500 (but preferably many more) for an independent text.

The tests rest upon stipulated contrasts between a base-set (the main sample of the current target-author) and a counter-set (comprising one or more of the remaining authorial samples). Table 2, a tiny Microsoft Excel worksheet, offers a concise but limited model of the procedure. The hierarchical array of word-types shown in Columns F and G runs down as usual from 'the' and 'and'. Their incidence in Waller, the base-set for this first series, is shown in Column I. In Columns A to E, the Waller set is broken into five segments of almost equal size, with the remainder added to segment five, increasing it from 2,767 to 2,770. The figures in Column H indicate how many of the five segments contain each word and run down, accordingly, from 5 to 1. This makes a simple measure of Waller's consistency in the use of each word-type in turn. (Even in the present list of extremely common words, 'your' appears in only three of the five segments. In the full list of 2,876 word-types, of course, 1 is much the most common count.) Column J and K treat of the chosen counter-set, the main set of Marvell. Column J shows which of the

Table 1 Eight seventeenth century poems. Results of the Delta tests on the 150 most common words of the main database

	Waller, Of the Danger 1322 Words Waller 1st: Marvell 11th			**Marvell, The First Anniversary 3131 Words Marvell 1st: Waller 10th**			**Waller, Instructions to a Painter 2605 Words Waller 1st: Marvell 7th**			**Marvell(?), Last Instructions 7693 Words Marvell 1st: Waller 16th**		
		Delta	**Delta z**		**Delta**	**Delta z**		**Delta**	**Delta z**		**Delta**	**Delta z**
1	Waller	1.549	−1.684	Marvell	1.058	3.517	Waller	1.278	−2.198	Marvell	1.094	−1.498
2	Prior	1.579	−1.492	Cowley	1.360	−1.263	Dryden	1.442	−1.300	Durfey	1.108	−1.416
3	Milton	1.584	−1.459	Prior	1.426	−0.773	Cowley	1.459	−1.205	Dryden	1.180	−0.982
4	Dryden	1.637	−1.114	Congreve	1.446	−0.624	Settle	1.462	−1.191	Prior	1.185	−0.949
5	Durfey	1.667	−0.916	Dryden	1.454	−0.563	Milton	1.528	−0.829	Tate	1.196	−0.883
6	Congreve	1.704	−0.676	Gould	1.478	−0.382	Tate	1.541	−0.758	Cowley	1.219	−0.744
7	Cowley	1.708	−0.654	Durfey	1.482	−0.356	Marvell	1.553	−0.694	Swift	1.221	−0.731
8	Tate	1.712	−0.628	Tate	1.487	−0.314	Durfey	1.598	−0.447	Settle	1.237	−0.639
9	Settle	1.727	−0.528	Denham	1.496	−0.251	Prior	1.599	−0.440	Milton	1.244	−0.594
10	Denham	1.730	−0.512	Waller	1.499	−0.224	Butler	1.616	−0.346	Congreve	1.248	−0.572
11	Marvell	1.762	−0.298	Settle	1.513	−0.122	Congreve	1.628	−0.281	Oldham	1.280	−0.380
12	Swift	1.766	−0.272	Milton	1.518	−0.083	Shadwell	1.650	−0.158	Butler	1.284	−0.353
13	Butler	1.793	−0.101	Oldham	1.539	0.069	Denham	1.669	−0.056	Shadwell	1.301	−0.251
14	Shadwell	1.804	−0.027	Butler	1.548	0.137	Gould	1.681	0.011	Gould	1.322	−0.125
15	Oldham	1.858	0.323	Cotton	1.561	0.232	Swift	1.686	0.037	Dorset	1.329	−0.083
16	Gould	1.873	0.421	Sedley	1.570	0.304	Oldham	1.688	0.050	Waller	1.347	0.025
17	Dorset	1.880	0.468	Swift	1.576	0.345	Sedley	1.802	0.674	Sedley	1.413	0.428
18	Cotton	1.904	0.624	Shadwell	1.581	0.386	Rochester	1.809	0.710	Radcliffe	1.456	0.686
19	Rochester	1.906	0.637	Dorset	1.586	0.421	Dorset	1.812	0.728	Denham	1.463	0.725
20	Sedley	1.945	0.888	Behn	1.649	0.896	Cotton	1.813	0.732	Cotton	1.464	0.735
21	Behn	1.959	0.982	Radcliffe	1.652	0.915	Behn	1.864	1.010	Behn	1.522	1.084
22	Wharton	1.967	1.035	Wharton	1.654	0.928	Wharton	1.881	1.106	Wharton	1.555	1.282
23	Phillips	2.025	1.408	Brome	1.667	1.024	Radcliffe	1.919	1.313	Brome	1.561	1.318
24	Radcliffe	2.030	1.445	Rochester	1.689	1.187	Brome	1.999	1.753	Rochester	1.576	1.406
25	Brome	2.136	2.132	Phillips	1.748	1.629	Phillips	2.004	1.781	Phillips	1.759	2.514

(Continued)

Table 1 Continued

	Second Advice 2867 Words Marvell 1st: Waller 20th			**Third Advice 3638 Words Marvell 1st: Waller 21st**			**Fourth Advice 1103 Words Waller 3rd: Marvell 14th**			**Fifth Advice 1209 Words Marvell 9th: Waller 23rd**		
		Delta	**Delta z**		**Delta**	**Delta z**		**Delta**	**Delta z**		**Delta**	**Delta z**
1	Marvell	1.253	−1.924	Marvell	1.161	−2.165	Gould	1.780	−1.659	Cotton	1.875	−1.247
2	Durfey	1.304	−1.488	Cowley	1.285	−1.186	Denham	1.800	−1.485	Durfey	1.884	−1.159
3	Prior	1.332	−1.248	Durfey	1.316	−0.940	Waller	1.819	−1.331	Radcliffe	1.896	−1.043
4	Cowley	1.360	−1.013	Tate	1.328	−0.845	Cowley	1.848	−1.079	Oldham	1.900	−0.994
5	Congreve	1.370	−0.922	Swift	1.345	−0.717	Tate	1.858	−0.992	Butler	1.907	−0.928
6	Butler	1.405	−0.625	Prior	1.348	−0.691	Butler	1.874	−0.856	Settle	1.909	−0.907
7	Oldham	1.411	−0.575	Oldham	1.349	−0.685	Oldham	1.877	−0.831	Gould	1.917	−0.824
8	Settle	1.425	−0.457	Gould	1.358	−0.610	Dryden	1.881	−0.799	Tate	1.928	−0.711
9	Dryden	1.425	−0.456	Dorset	1.386	−0.388	Settle	1.908	−0.572	Marvell	1.938	−0.609
10	Denham	1.426	−0.452	Butler	1.396	−0.312	Shadwell	1.924	−0.437	Sedley	1.956	−0.433
11	Tate	1.439	−0.338	Sedley	1.398	−0.299	Swift	1.948	−0.232	Brome	1.959	−0.395
12	Swift	1.453	−0.221	Settle	1.401	−0.273	Durfey	1.952	−0.193	Dorset	1.960	−0.393
13	Gould	1.458	−0.173	Dryden	1.402	−0.268	Cotton	1.958	−0.142	Milton	1.969	−0.300
14	Milton	1.465	−0.121	Milton	1.410	−0.204	Marvell	1.974	−0.006	Cowley	1.971	−0.281
15	Shadwell	1.466	−0.111	Cotton	1.418	−0.139	Brome	2.016	0.343	Swift	1.974	−0.249
16	Dorset	1.490	0.092	Congreve	1.447	0.087	Congreve	2.026	0.432	Behn	2.014	0.161
17	Sedley	1.496	0.144	Shadwell	1.468	0.251	Radcliffe	2.041	0.559	Congreve	2.058	0.610
18	Cotton	1.567	0.755	Denham	1.497	0.481	Dorset	2.058	0.708	Dryden	2.065	0.680
19	Brome	1.576	0.831	Radcliffe	1.498	0.487	Phillips	2.060	0.720	Shadwell	2.073	0.755
20	Waller	1.605	1.073	Brome	1.525	0.701	Milton	2.085	0.933	Prior	2.073	0.759
21	Behn	1.627	1.261	Waller	1.578	1.116	Prior	2.091	0.988	Denham	2.078	0.813
22	Phillips	1.627	1.263	Behn	1.595	1.252	Behn	2.116	1.195	Rochester	2.093	0.965
23	Rochester	1.637	1.346	Rochester	1.604	1.323	Sedley	2.140	1.404	Waller	2.129	1.323
24	Radcliffe	1.671	1.641	Phillips	1.647	1.667	Wharton	2.166	1.621	Wharton	2.212	2.169
25	Wharton	1.681	1.719	Wharton	1.735	2.357	Rochester	2.176	1.710	Phillips	2.219	2.237

Table 2 Waller and Marvell. Outline of the new procedure

	Segments of Base-set							Base-set		Counterset		Trial-pieces		Test-pieces	
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Waller1	Waller2	Waller3	Waller4	Waller5	WORDS		Waller	Waller	Marvell	Marvell	Danger	FirstAnn	Painter	LastInst
2	2767	2767	2767	2767	2770				13838		20151	1322	3131	2605	7693
3								COUNT	SUM	COUNT	SUM				
4	20	20	20	19	19				20	20	20	19	20	20	20
5	2767	2767	2767	2767	2770				13838		20151	1322	3131	2605	7693
6	7.23	7.23	7.23	6.87	6.86				1.45		0.99	14.37	6.39	7.68	2.60
7															
8	801	743	740	792	740				3816		4799	361	826	781	2060
9	2767	2767	2767	2767	2770				13838		20151	1322	3131	2605	7693
10	289.48	268.52	267.44	286.23	267.15				275.76		238.15	273.07	263.81	299.81	267.78
11															
12	153	169	147	132	123	1	the	5	724	1	1024	96	199	188	451
13	100	98	91	96	74	2	and	5	459	1	704	43	125	109	331
14	56	57	44	52	52	3	of	5	261	1	295	47	55	45	107
15	50	25	42	49	38	4	to(inf.)	5	204	1	263	12	23	51	114
16	23	44	31	69	35	5	his	5	202	1	194	31	30	61	146
17	35	36	33	54	37	6	a	5	195	1	269	13	58	34	94
18	34	29	45	29	40	7	in(p)	5	177	1	293	15	32	32	124
19	48	17	32	32	37	8	our	5	166	1	70	9	13	18	39
20	45	30	19	37	23	9	with	5	154	1	205	20	39	32	101
21	28	24	37	26	29	10	to(p)	5	144	1	170	13	23	16	66
22	21	23	44	25	28	11	as	5	141	1	198	8	24	17	39
23	41	29	21	27	18	12	their	5	136	1	140	18	38	73	64
24	18	12	28	28	46	13	we	5	132	1	67	0	14	8	7
25	13	18	18	33	40	14	he	5	122	1	192	6	32	18	90
26	17	13	25	27	29	15	but	5	111	1	228	5	31	15	79
27	28	29	10	17	21	16	from	5	105	1	71	8	16	20	38
28	19	21	15	22	23	17	all	5	100	1	153	6	21	12	74
29	33	36	31	0	0	18	your	3	100	1	48	1	4	5	3
30	20	17	8	21	27	19	that(rp)	5	93	1	116	6	16	6	38
31	19	16	19	16	20	20	they	5	90	1	99	4	33	21	55

NB: (inf) = infinitive particle; (p) = preposition; (rp) = relative pronoun.

word-types actually occur in Column K, the list of scores for Marvell's main set. (Registering only 1 here and only 1 or 0 in the full list, Column J serves little purpose until we turn to multi-author counter-sets.) The remaining four columns, L to O, show corresponding scores for four of the chosen texts. The separation between 'trial-pieces' and 'test-pieces' allows the specimens actually at issue to be held aside until the desired trial-runs are complete and a set of stipulations appropriate to the particular case has been determined. It is good practice, obviously, not even to introduce the test-pieces until that stage of proceedings. In the present article, however, the distinction between 'trial-pieces' and 'test-pieces' is inoperative because the same few stipulations will be used throughout.

The overall results are shown in Rows 4 to 6 and 8 to 10. Rows 4 to 6 treat word-types. Row 4 shows the sum of occurrences in each text-specimen of whatever word-types are listed in column G. Row 5 shows the length of each specimen. Row 6 registers the number of word-types as a rate per thousand of the corresponding entry in Row 5. Rows 8 to 10 treat the sum-total of word-tokens in the same fashion. In column D, for example, 19 of the 20 types shown in column G occur in this segment; and these 19 types occur a total of 792 times, which is a rate of 286.23 per thousand.

With texts of uniform length—segments of plays or novels, for example—the relative occurrence-rates of the word-types are a valuable differentia.[7] But, as has often been observed, the type/token ratio is much affected by text-length. With texts of different length, like a group of poems, the occurrence-rate of the word-tokens themselves is a more useful measure. Even in the miniature example of Table 2, the occurrence-rates of the three Waller entries exceed those for the three associated with Marvell. But the gap between the lowest Waller and the highest Marvell—273.07 against 267.78 per thousand—is too narrow to be convincing. A more stringent approach is needed.

The stipulated contrasts between base-set and counter-set are the heart of the affair. In the simple form shown in Table 2, the only operative criterion is reflected in the descending order of Waller's scores in Column I. The fact that this is Waller's word-list, not Marvell's, is evident in the many perturbations of rank-order in Column K. But the comparatively weak distinctions shown in Row 10 need to be strengthened and controlled.

If all the data in the full work-sheet for Waller, from Row 12 down to Row 2,887, are sorted on the basis of the counts in Column H, it is possible to discard all those word-types that do not meet a stipulated level of consistency across the five segments of the main Waller-set as represented in Columns A to E. If the remaining data are then sorted on the basis of Column K, it is possible to discard all those word-types that exceed a specified level of frequency in Marvell. If the stipulations are too weak, no consistent authorial difference will emerge. But if they are too strict, the surviving word-list can be too impoverished to yield a reliable outcome.

Table 3 shows the outcome of tests in which Waller and Marvell are opposed. In the top half of the table, Waller provides the base-set, Marvell the counter-set. In the bottom half, their roles are reversed. In each half of the table, we open the series with a simple overview. Beyond the fact that these are the words of one or other of the two opposed authorial sets, no further stipulation is imposed. In both overviews, the occurrence-rate for the chosen base-set is, by definition, 1,000 per 1,000. The occurrence-rate for each of the counter-sets and for such other specimens as may be introduced will never reach that level. In practice, an occurrence-rate between 750 and 850 per 1,000 is usual. Consider the present case, beginning with Overview A. Among the eight specimen-texts, the two by Waller show higher occurrence-rates than the four associated with Marvell. But the 'Fourth Advice' breaks the pattern, with an occurrence-rate higher than that for Waller's 'Of the Danger'. In Overview B, Marvell's 'First Anniversary' shows the highest score of the eight but the pattern of the other seven scores is not authorial. These overviews, then, do not support the idea that a given writer's overall word-list might serve as it stands as an accurate authorial discriminator.

The main difficulty here is that any truly idio-syncratic features of the two authorial frequency-lists are buried among the high frequencies of words

Table 3 Waller and Marvell. Two tests on eight poems

Waller versus Marvell		**Length**	**Overview A** **Base-set: No Stipulations** **Counter-Set: Ditto**			**Test 2A: Zeta test** **Base-set: >2ex5 Segs.** **Counter-set: Frequency <3**			**Test 3A: Iota test** **Base-set: <3ex5 Segs.** **Counter-set: Frequency Zero**		
			Types	**Tokens**	**per 1000**	**Types**	**Tokens**	**per1000**	**Types**	**Tokens**	**per 1000**
Waller	Base-set	13838	2876	13838	1000.00	185	967	69.88	1169	1494	107.96
Marvell	Counter-set	20151	1644	20151	798.37	122	179	8.88	0	0	0.00
Free agents											
Waller	Of the Danger	1322	464	1322	823.75	44	62	46.90	57	65	49.17
Marvell	First Anniv	3131	715	3131	805.17	50	66	21.08	72	76	24.27
Waller	Inst. Painter	2605	702	2605	864.11	72	122	46.83	106	143	54.89
Marvell?	Last Inst.	7693	1091	7693	753.15	76	159	20.67	164	216	28.08
Marvell??	2nd Advice	2867	632	2867	763.86	32	47	16.39	71	87	30.35
Marvell??	3rd Advice	3638	695	3638	758.66	47	61	16.77	74	90	24.74
??	4th Advice	1103	361	1103	841.34	33	41	37.17	36	43	38.98
??	5th Advice	1209	331	1209	775.85	27	31	25.64	22	24	19.85
Marvell versus Waller		**Length**	**Overview B** **Base-set: No Stipulations** **Counter-set: Ditto**			**Test 2B: Zeta test** **Base-set: >2ex5 Segs.** **Counter-set: Frequency <3**			**Test 3B: Iota test** **Base-set: <3ex5 Segs.** **Counter-set: Frequency Zero**		
			Types	**Tokens**	**per 1000**	**Types**	**Tokens**	**per 1000**	**Types**	**Tokens**	**per 1000**
Marvell	Base-set	20151	4323	20151	1000.00	354	2143	106.35	2521	3216	159.60
Waller	Counter-set	13838	1644	12041	870.14	196	278	20.09	0	0	0.00
Free agents											
Waller	Of the Danger	1322	470	1090	824.51	53	60	45.39	58	63	47.66
Marvell	First Anniv	3131	835	2687	858.19	135	236	75.38	160	191	61.00
Waller	Inst. Painter	2605	681	2197	843.38	65	88	33.78	85	100	38.39
Marvell?	Last Inst.	7693	1360	6198	805.67	185	417	54.21	381	514	66.81
Marvell??	2nd Advice	2867	721	2294	800.14	99	151	52.67	141	163	56.85
Marvell??	3rd Advice	3638	839	2962	814.18	130	230	63.22	184	232	63.77
??	4th Advice	1103	362	926	839.53	30	34	30.83	42	48	43.52
??	5th Advice	1209	367	987	816.38	39	59	48.80	56	71	58.73

that everybody uses. As the little array of raw data in Table 2 reveals, the twenty most common word-types of Waller's set amount to about a quarter of all the word-tokens in each of the specimens. The steepness and the uniformity of the descending hierarchy were observed by George K. Zipf (1949) and enshrined in 'Zipf's Law'. But as a recent article (Hoover, 2004) has shown, the Delta procedure continues to be effective for at least the top 600 word-types of a well-founded word-list. A merely additive procedure, however, like that used for Overviews A and B, is too crude for the task in hand.

A second difficulty, I believe, affects the use of any measure of divergence for the analysis of complete frequency hierarchies. It stems not from the tyranny of the large numbers at the head of the list but from subtler arithmetical differences among the data at different levels of a given frequency-table. In the topmost stratum, robust statistical comparisons can be derived from differently sized divergences from a mean or median score. For each common word-type in a range of specimens, a bell-shaped curve provides a sound footing for a range of useful operations. These curves are always skewed to the right or positive side because such data yield zeroes and very low scores for particular word-types more often than very high scores. Sentences of some length where 'the' does not occur are more common than those where it occurs in abundance. But the upward range is the more extensive. It is not hard, for instance, to devise meaningful sentences where 'the' makes up a quarter or more of all the word tokens: 'The bald Anglo the blonde had eyed the night before lay on the dirty mat. The haft of the Greek's dagger stood in the hairy chest. The blind flapped and the scattered papers stirred in the breeze from the broken pane. The blonde officer sighed at the enormity of the task ahead. The upside? Yeah, right—the Anglo sure was off the hot-list.' Although such levels cannot sensibly be sustained, many a dreary, over-circumstantial novel suffers in the attempt.

In the next stratum from the top, zeroes and singletons are common but are not infrequently matched against scores that run up from a handful into double figures. In the lowest and most extensive stratum, zeroes by far predominate but are usually matched against scattered singletons in simple binary contrasts. Means and standard deviations can certainly be extracted from the frequency-patterns of these lower strata. To put them to the same uses as those of the top stratum seems inappropriate in principle and can yield disconcerting results. Let us return to Table 3 and consider some examples of tests where the two lower strata are separated from the top one and also from each other. I have been labelling them as 'Zeta' and 'Iota' tests as a matter of convenience and as a way of emphasizing that Omega, the last word on these matters, is not yet within reach.

In the Zeta tests 2A and 2B ('Delta' being Test 1), two stipulations are imposed. After breaking each base-set, in turn, into five equal segments, we retain only those word-types that occur in at least three of the five. Of these, we discard those that occur fewer than three times in the counter-set. In Test 2A, Waller's list of word-types is reduced from 2,876 to 185. In Test 2B, Marvell's list of 4,323 is reduced to 354. As stipulated, the contrast between each base-set and the corresponding counter-set is very marked. The real point of interest, however, lies in the behaviour of the eight independent specimens. In Test 2A, the occurrence-rates for the two Waller entries, at 46.90 and 46.83 per 1,000, comfortably exceed all their rivals. In Test 2B, the occurrence-rates for the four poems associated with Marvell exceed the other four.

The Iota tests 3A and 3B focus on the lowest stratum of the word-lists. The first stipulation is that a word-type should not appear in more than two of the five segments of the appropriate base-set. The second is that it should not occur at all in the corresponding counter-set. The Waller-list is thus reduced to 1,169 uncommon word-types, the Marvell to 2,521. In Test 3A, the scores for the two Waller entries far exceed their rivals. In Test 3B, three of the texts associated with Marvell exceed the rest. But the authorial pattern is broken by the 'Fifth Advice', which scores a trifle higher than the 'Second Advice'. The range of the eight scores indicates that this breach is due to a higher than expected score for the 'Fifth Advice'. The 'Second Advice' does not fall far short of its true partners.

The score for the 'Fifth Advice' makes a useful point. It reflects the fact that the tests in Table 3 are specifically designed to distinguish Waller and Marvell from each other and not from anybody else. A text by any other poet is not governed by such differentiae and remains a wild-card, free to include words from the base-set and not *necessarily* much affected by the exclusion of words from the counter-set. The stipulated contrasts may operate successfully on such texts but cannot be expected to do so. It is as if, after establishing a set of differentiae to distinguish chalk from cheese, one tried them on a specimen of bone. The likeness to chalk in hardness, density, and colour is adventitious. This is not chalk at all. A more broadly based set of differentiae is required for the more general task.

4.2 One against many: Waller and twenty-four others

Simplicity of exposition has its place in a demonstration-piece. In practice, however, head to head contests between two putative authors are best reserved for situations where no other candidates need be considered. That situation can arise from outside knowledge, as when only John Dryden and the Earl of Mulgrave need be considered as possible authors of the celebrated 'Essay on Satire'. It can arise, by experiment, from the use of exploratory tests like principal component analysis and the Delta procedure. But, as we have just seen, a premature head to head contest can easily go awry.

Once a putative author has been identified, by whatever means, it is possible to verify his or her claim by using variant forms of the Zeta and Iota tests. This can be done by matching a chosen authorial base-set, as before, against a multi-author counter-set. A table corresponding to Table 2 is established. Columns A to I are retained unaltered. But Column K is now occupied by the sum of all the scores, for each successive word-type, of as many other authorial sets as may be desired. Each of those sets occupies its own column and contributes to the sum. And Column J is now occupied by counts of the number of those authors who use each of the word-types. Any stipulated contrast between base-set and counter-set can now be applied to whatever specimens may seem appropriate.

In the present case, Waller supplies the initial base-set. The counter-set is composed of the twenty-four other authorial sets, Marvell among them, that make up our main database. The specimens for examination are the same eight poems as before. They are tested against the twenty-four separate authorial sets; twenty-one independent texts by those same poets; and twenty texts by other Restoration poets.

In Table 4, the Zeta test is based, as before, on two stipulations. The first excludes all those word-types that occur in fewer than three of Waller's five segments. The second excludes those word-types that occur in more than twenty-two of the twenty-four other authorial sets. The second stipulation serves not only to help establish a firm contrast between Waller and the rest but also to exclude those very common words that are used in other statistical procedures. The effect of the two stipulations is to reduce Waller's 2,876 word-types to 259. These yield him 1,319 word-tokens at a rate of 95.32 per 1,000. The half-million words of the counter-set yield 16,517 word-tokens at a rate of only 31.72 per 1,000.

In the top central panel of Table 4, it can be seen that, at 72.94 and 59.00, the scores for the two Waller poems easily exceed those for the other six. The scores for Group A, in the panel below it, show that the twenty-four authorial components of the counter-set all fall well short of the two Waller poems but that some of them exceed the least Waller-like of the other six specimens. The mean of these twenty-four sets is 31.24 per 1,000 and the highest of them is only 41.84. The low standard deviation of 6.06 suggests that the stipulations employed have fallen with rather an even hand across this group and also helps to emphasize the sharp divergence of the Waller pieces from the rest.

But the true force of the test is felt in Groups B and C, where the scores for the independent specimens are arrayed. These two lowest sections of the central panel show that the scores for the two Waller poems are still not matched by any of these forty-one texts. Neither those by members of the set nor those by outsiders produce a serious rival. Two of the pieces by outsiders score over 50 per 1,000. They are Blackmore's *King Arthur* and Flatman's

Table 4 Waller and 24 others. Two tests on eight poems: Waller's world-list

		Length	Zeta test Base-set: >2ex5 Segs. Counter-set: <23 Poets				Iota test Base-set: <3ex5 Segs. Counter-set: <11 Poets			
			Types	Tokens	per 1000	Rank	Types	Tokens	per 1000	Rank
Waller	Base-set	13838	259	1319	95.32		1127	1380	99.73	
24 poets	Counter-set	520672	4350	16517	31.72		5485	8406	16.14	
Test-pieces										
Waller	Of the Danger	1322	59	78	59.00	2	39	46	34.80	2
Marvell	First Anniv	3131	81	116	37.05		66	68	21.72	
Waller	Inst. Painter	2605	105	190	72.94	1	81	103	39.54	1
Marvell?	Last Inst.	7693	126	272	35.36		125	158	20.54	
Marvell??	2nd Advice	2867	60	110	38.37		66	74	25.81	
Marvell??	3rd Advice	3638	79	124	34.08		59	67	18.42	
??	4th Advice	1103	43	50	45.33	3	21	23	20.85	
??	5th Advice	1209	30	36	29.78		10	10	8.27	
Group A. Components of counter-set										
MAX					41.84				21.05	
MEAN					31.24				15.84	
ST DEV					6.06				2.85	
Behn		21705	173	672	30.96	13	167	290	13.36	19
Brome		29539	182	694	23.49	22	218	318	10.77	24
Butler		30932	178	792	25.60	19	275	441	14.26	14
Congreve		30917	218	1067	34.51	9	340	633	20.47	2
Cotton		12625	144	321	25.43	20	146	165	13.07	21
Cowley		19272	196	710	36.84	5	206	321	16.66	10
Denham		30094	220	1024	34.03	10	324	547	18.18	6
Dorset		9586	126	233	24.31	21	113	129	13.46	18
Dryden		18238	201	738	40.46	3	236	326	17.87	8
Durfey		18757	186	671	35.77	6	223	320	17.06	9
Gould		29110	201	832	28.58	14	281	472	16.21	11
MarvellB		20151	198	625	31.02	12	257	381	18.91	4
Milton		18924	170	514	27.16	17	234	340	17.97	7
Oldham		32462	206	906	27.91	15	324	490	15.09	13
Phillips		29004	206	1032	35.58	7	261	411	14.17	16
Prior		32000	228	1339	41.84	1	362	655	20.47	3
Radcliffe		11889	119	253	21.28	24	126	156	13.12	20
Rochester		12725	128	297	23.34	23	134	160	12.57	23
Sedley		10304	125	284	27.56	16	115	131	12.71	22
Settle		24080	208	989	41.07	2	270	443	18.40	5
Shadwell		14540	178	516	35.49	8	173	232	15.96	12
Swift		30974	215	833	26.89	18	295	440	14.21	15
Tate		20333	194	758	37.28	4	276	428	21.05	1
Wharton		12511	150	417	33.33	11	129	177	14.15	17
Group B. Independent texts by these poets										
MAX					36.99				28.25	
MEAN					27.92				18.31	
ST DEV					6.67				5.41	
Behn16	Isle of Love	16419	164	550	33.50	6	176	298	18.15	11
Brome16	Answer	1385	19	20	14.44	21	22	22	15.88	16

(Continued)

Table 4 Continued

		Length	Zeta test Base-set: >2ex5 segs. Counter-set: <23 poets				Iota test Base-set: <3ex5 segs. Counter-set: <11 poets			
			Types	Tokens	per1000	Rank	Types	Tokens	per1000	Rank
Butler01	Plagiaries	1217	26	34	27.94	12	21	24	19.72	10
Butler02	Weakness	1401	25	30	21.41	17	15	15	10.71	19
Congreve14	Imposs. Thing	1376	28	34	24.71	16	23	24	17.44	12
Cotton09	Ireland2	1830	28	39	21.31	18	14	15	8.20	21
Cotton11	Epistle	1537	25	26	16.92	20	13	15	9.76	20
Cowley17	Royal Society	1315	32	41	31.18	9	27	28	21.29	8
Cowley18	Davideis2	6812	118	252	36.99	1	106	147	21.58	6
Dryden03	Absalom	7824	122	283	36.17	3	140	221	28.25	1
Dryden02	Epistle15	1650	45	61	36.97	2	35	41	24.85	2
Durfey	Malecontent	7817	114	250	31.98	7	106	134	17.14	14
Gould13	BeauxEsprits	6020	100	177	29.40	10	70	79	13.12	17
Gould16	Fanaticism	1995	57	67	33.58	4	31	33	16.54	15
Milton02	PR	15694	166	527	33.58	5	216	332	21.15	9
Milton03	SA	12885	133	356	27.63	13	181	275	21.34	7
Oldham07	Ona Woman	1298	18	23	17.72	19	15	15	11.56	18
Oldham13	Ben Johnson	2237	43	57	25.48	14	48	49	21.90	5
Phillips	Lady E.C.	1210	28	30	24.79	15	26	29	23.97	4
Prior18	Henry&Emma	6033	96	191	31.66	8	97	149	24.70	3
Swift18	Verses	3206	57	93	29.01	11	46	55	17.16	13
Group C. Independent texts by other poets										
MAX					52.58				29.68	
MEAN					36.61				19.97	
ST DEV					8.21				4.77	
Baker	Death	1077	24	26	24.14	20	13	13	12.07	20
Blackmore	K.Arthur	6986	127	352	50.39	2	125	203	29.06	2
Caryll	Naboth	3892	69	119	30.58	16	56	69	17.73	14
Chamb'l'ne	Pharonn	3617	74	114	31.52	15	74	90	24.88	3
Chudleigh	Gloucester	2743	74	109	39.74	7	39	49	17.86	12
Davenant	Gondibert	5167	97	198	38.32	8	86	100	19.35	10
Duke	Paris	3486	73	121	34.71	11	48	63	18.07	11
Fane	Pindar	1909	62	93	48.72	3	30	32	16.76	16
Flatman	Rupert	1065	50	56	52.58	1	23	23	21.60	7
Garth	Dispens5	2704	59	88	32.54	14	48	54	19.97	9
Heyrick	Atlantis	8797	150	364	41.38	6	143	185	21.03	8
Hutchinson	Order	6792	108	227	33.42	12	135	168	24.73	4
Mulgrave	Poetry	2765	71	103	37.25	9	38	40	14.47	18
Norris	Passion	1319	36	46	34.87	10	28	32	24.26	5
Pordage	Med. Rev.	3103	52	103	33.19	13	29	44	14.18	19
Sprat	Oliver	2522	73	114	45.20	4	42	45	17.84	13
Thompson	Mid. Moon	3250	66	87	26.77	18	49	51	15.69	17
Tutchin	Honesty	2287	46	64	27.98	17	46	52	22.74	6
Wase	Divination	2156	70	92	42.67	5	59	64	29.68	1
Wild	Iter Boreale	3321	62	87	26.20	19	49	58	17.46	15

pindaric ode 'On the Death of the Illustrious Prince Rupert'.

Both poems share in the nationalistic rodomontade in which Waller is among the more extravagant of his generation. He is a strong exponent of attitudes that (springing from a narrower tribalism) were gaining ground in Europe, were later to cross the Atlantic, and (chiefly to the detriment of humanity at large) were to influence European and world history in the ensuing 300 years.

> Sing we the Glory of triumphant Arms.
> So shall all Tyrants yield.
> May the like Fortune meet all those
> Who vainly dare oppose
> Our Monarch's sacred Law,
> Our Nation's noble Rage.

This fragment is from a pindaric ode 'On our late Victories by Land and Sea'. In admitting authorship (while disavowing the bombastic sentiments), I note that it is studded with words which Waller uses to like purposes and which help to distinguish him from most others. Of Waller's 2,876 word-types, 259 meet the stipulations stated above. The ten most common of them are: sacred, rage, glory, sing, law, fortune, equal, reign, yield, triumph. The most common such function-word is 'like', used as an adjective.

The effectiveness of the Zeta test on this occasion is undeniable. Its more general reliability remains in doubt. A set of stipulations that identifies texts characteristic of Waller's maturity yields a word-list that might well be less accurate in identifying the love-songs of his youth or the immense religious poems of his dotage. But such words as the adjective 'like' might persist throughout his long career.

The right-hand panels of Table 4 show corresponding results for the Iota test. The stipulations employed on this occasion exclude all those words that occur in more than two of Waller's five segments and all those that occur in more than ten of the twenty-four authorial components of the counter-set. Waller's 2,876 word-types are reduced to 1,127. The 1,380 word-tokens occur at a rate of 99.73 per 1,000 in the base-set. They are matched by 8,406 at a rate of 16.14 in the counter-set.

At 34.80 and 39.54 per 1,000, the two Waller poems easily outscore the other six. They leave the members of Group A far behind though the highest of these lie above some of the six non-Waller pieces. Some of the independent texts in Groups B and C score higher than any in Group A. But none of them approach Waller's pair. The standard deviations for all three groups are low.

Of the individual texts, the least remote are *Absalom and Achitophel* and Christopher Wase's *Divination*. The former is a celebrated political satire, the latter a pro-Wallerian contribution to the debate focussed on his 'Instructions to a Painter'. In this less common range of words, we are obviously tapping a different vein from that of Waller's patriotic effusions. The word-list now opens with the following ten word-types: portion, armies, exceed, fishes, indite, maker, Christians, tragedy, Chloris, Turks. These and their successors have little in common save for the salient point: they are words used by Waller but not by most of his contemporaries. It is worth noticing, moreover, that they are not truly rare words. There are few rarities even among the 127 word-types used by Waller alone. None of them occur more than twice in Waller but the first six appear in two of his segments. The list begins: repressed, Isaiah, Pandora, daw, piracy, displaying, vizier, enlargement, Antarctic, forebodes.

4.3 One against many: Marvell and twenty-four others

The two tests whose results are set out in Table 5 have Marvell as base-set and the other twenty-four poets as counter-set. The size of the counter-set is altered by the replacement of Marvell by Waller among the twenty-four. The stipulations are exact Marvellian equivalents of those used for Table 4.

As the right-hand panel of Table 5 makes clear, Marvell responds almost as well as Waller to the Iota test. Yet the 2,344 word-types that satisfy our stipulations are miscellaneous indeed. As with the corresponding set in Waller, few of them are rare. The upper range of the list, the words that Marvell shares with ten of the other poets, is marked by the language of pastoral. But, as it must, the list grows more idiosyncratic as more of the other poets are excluded. The set of words used by Marvell alone

Table 5 Marvell and twenty four others. Two tests on eight poems: Marvell's word-list

		Length	Zeta test Base-set: >2ex5 Segs. Counter-set: <23 Poets				Iota test Base-set: <3ex5 Segs. Counter-set: <11 Poets			
			Types	Tokens	per 1000	Rank	Types	Tokens	per 1000	Rank
Marvell	Base-set	20151	389	2033	100.89		2344	2879	142.87	
24 poets	Counter-set	514359	6473	24216	47.08		8754	13091	25.45	
Test-pieces										
Waller	Of the Danger	1322	68	84	**63.54**	**2**	46	48	36.31	
Marvell	First Anniv	3131	137	208	**66.43**	**1**	129	147	**46.95**	**2**
Waller	Inst. Painter	2605	88	139	53.36	**4**	68	77	29.56	
Marvell?	Last Inst.	7693	197	384	49.92	**5**	282	365	**47.45**	**1**
Marvell??	2nd Advice	2867	101	156	**54.41**	**3**	117	131	**45.69**	**3**
Marvell??	3rd Advice	3638	121	172	47.28	**6**	125	143	**39.31**	**4**
??	4th Advice	1103	33	39	35.36		27	29	26.29	
??	5th Advice	1209	37	48	39.70		27	31	25.64	
Group A. Components of counter-set										
MAX					**64.15**				**38.68**	
MEAN					**46.91**				**25.29**	
ST DEV					**7.94**				**5.13**	
Behn		21705	272	1248	**57.50**	**3**	306	516	23.77	16
Brome		29539	261	996	33.72	24	360	519	17.57	23
Butler		30932	270	1206	38.99	22	422	725	23.44	17
Congreve		30917	343	1877	**60.71**	**2**	549	926	**29.95**	**3**
Cotton		12625	238	591	46.81	13	277	335	26.53	9
Cowley		19272	293	1074	55.73	4	337	498	25.84	10
Denham		30094	311	1484	49.31	9	531	814	27.05	7
Dorset		9586	175	344	35.89	23	155	194	20.24	21
Dryden		18238	285	930	50.99	7	375	520	28.51	4
Durfey		18757	269	782	41.69	16	363	501	26.71	8
Gould		29110	286	1139	39.13	21	427	650	22.33	18
Milton		18924	303	1214	**64.15**	**1**	494	732	**38.68**	**1**
Oldham		32462	305	1437	44.27	14	503	775	23.87	15
Phillips		29004	312	1419	48.92	11	404	607	20.93	20
Prior		32000	308	1567	48.97	10	519	881	27.53	6
Radcliffe		11889	203	479	40.29	19	234	294	24.73	13
Rochester		12725	230	503	39.53	20	203	256	20.12	22
Sedley		10304	213	428	41.54	17	203	247	23.97	14
Settle		24080	293	1149	47.72	12	438	675	28.03	5
Shadwell		14540	237	589	40.51	18	254	322	22.15	19
Swift		30974	315	1321	42.65	15	508	782	25.25	12
Tate		20333	285	1071	52.67	6	487	756	**37.18**	**2**
Waller		13838	240	684	49.43	8	254	357	25.80	11
Wharton		12511	226	684	54.67	5	151	209	16.71	24
Group B. Independent texts by these poets										
MAX					**61.00**				**40.70**	
MEAN					**41.83**				**27.16**	
ST DEV					**9.12**				**5.94**	
Behn16	Isle of Love	16419	231	979	**59.63**	**2**	275	477	29.05	9
Brome16	Answer	1385	38	54	38.99	15	28	33	23.83	16
Butler01	Plagiaries	1217	29	33	27.12	21	34	36	29.58	8

(Continued)

Table 5 Continued

		Length	Zeta test Base-set: >2ex5 Segs. Counter-set: <23 Poets				Iota test Base-set: <3ex5 Segs. Counter-set: <11 Poets			
			Types	Tokens	per1000	Rank	Types	Tokens	per1000	Rank
Butler02	Weakness	1401	36	44	31.41	19	36	38	27.12	12
Congreve14	Imposs. Thing	1376	44	54	39.24	14	52	56	**40.70**	**1**
Cotton09	Ireland2	1830	51	78	42.62	9	39	49	26.78	13
Cotton11	Epistle	1537	35	44	28.63	20	26	27	17.57	20
Cowley17	Royal Society	1315	48	61	46.39	6	37	39	29.66	6
Cowley18	Davideis2	6812	199	380	**55.78**	**3**	165	209	30.68	5
Dryden03	Absalom	7824	160	308	39.37	13	176	232	29.65	7
Dryden02	Epistle15	1650	42	57	34.55	17	57	62	**37.58**	**2**
Durfey	Malecontent	7817	182	341	43.62	8	162	191	24.43	15
Gould13	BeauxEsprits	6020	128	241	40.03	12	118	123	20.43	18
Gould16	Fanaticism	1995	62	80	40.10	11	37	40	20.05	19
Milton02	PR	15694	231	732	46.64	5	341	500	31.86	4
Milton03	SA	12885	193	605	46.95	4	281	412	**31.98**	**3**
Oldham07	Ona Woman	1298	46	59	45.45	7	28	28	21.57	17
Oldham13	Ben Johnson	2237	72	94	42.02	10	57	58	25.93	14
Phillips	Lady E.C.	1210	36	44	36.36	16	20	21	17.36	21
Prior18	Henry&Emma	6033	164	368	**61.00**	**1**	129	164	27.18	11
Swift18	Verses	3206	74	104	32.44	18	62	88	27.45	10
Group C. Independent texts by other poets										
MAX					**65.62**				**44.60**	
MEAN					**48.35**				**30.17**	
ST DEV					**13.72**				**7.65**	
Baker	Death	1077	42	49	45.50	16	28	29	26.93	14
Blackmore	K.Arthur	6986	160	357	51.10	10	171	251	**35.93**	**3**
Caryll	Naboth	3892	91	138	1.00	20	110	129	33.14	10
Chamb'l'ne	Pharonn	3617	123	202	55.85	5	103	116	32.07	12
Chudleigh	Gloucester	2743	107	180	**65.62**	**1**	45	47	17.13	19
Davenant	Gondibert	5167	145	244	47.22	14	133	184	35.61	6
Duke	Paris	3486	128	220	**63.11**	**2**	65	75	21.51	17
Fane	Pindar	1909	74	105	55.00	6	50	63	33.00	11
Flatman	Rupert	1065	46	54	50.70	11	37	38	35.68	5
Garth	Dispens5	2704	93	131	48.45	12	96	104	**38.46**	**2**
Heyrick	Atlantis	8797	209	537	**61.04**	**3**	229	297	33.76	8
Hutchinson	Order	6792	168	353	51.97	8	196	243	35.78	4
Mulgrave	Poetry	2765	82	111	40.14	18	44	52	18.81	18
Norris	Passion	1319	63	77	58.38	4	44	44	33.36	9
Pordage	Med. Rev.	3103	82	130	41.89	17	36	52	16.76	20
Sprat	Oliver	2522	88	133	52.74	7	55	59	23.39	16
Thompson	Mid. Moon	3250	84	103	31.69	19	78	80	24.62	15
Tutchin	Honesty	2287	76	109	47.66	13	90	102	**44.60**	**1**
Wase	Divination	2156	81	111	51.48	9	72	75	34.79	7
Wild	Iter Boreale	3321	103	154	46.37	15	78	93	28.00	13

opens with: ungirt, practising, departure, ambergris, Thwaites, tulip, tinkling, melons, perpetration, architects. While this little sample is not an encouraging beginning, the full list of 2,344 allows a cumulative effect strong enough to identify Marvell's work with almost perfect accuracy.

Of our first eight specimens, those associated with Marvell range down from 47.45 to 39.31 per

1,000 and so outscore the other four. They also outscore the twenty-four authorial sets of Group A. Of the twenty-one specimens in Group B, Congreve's comic tale, 'An Impossible Thing', at 40.70 per 1,000, outscores the 'Third Advice'. Of the twenty poems in Group C, John Tutchin's satire 'A Search after Honesty', at 44.60 per 1,000, also outscores the 'Third Advice'. The other three Marvell texts outscore all others.

Tutchin's brand of rough satire often yields unexpected resemblances to satirical poems by other authors. Congreve, like Milton, has a richer vocabulary than most even when his large authorial set is cut down. One effect is that he often uses more of the uncommon words than most and, accordingly, tends to encroach on the word-lists of his fellows. Since Tutchin was a child of about six and Congreve was yet unborn when the 'Third Advice' was first published, neither is a candidate for its authorship.

To resolve the question without benefit of such external evidence, the current set of stipulations can be modified. In the present instance, perfect accuracy is obtainable by discarding all the words that are used by more than four of the other twenty-four poets. The other obvious course is to set up one-on-one contests (as illustrated earlier) between the claimants. When the words they share are excluded by the usual sort of stipulations, the remainder afford a basis for weighing up their respective claims. I have yet to encounter a case where the Zeta and Iota tests fail when they are used in a genuine one-on-one end game.

The central panel of Table 5 shows the results of a Zeta test in which Marvell's base-set is matched against the counter-set of twenty-four authors. As with Waller, the exclusion of words that occur in fewer than three segments of the base-set and in more than twenty-two members of the counter-set yields a solid foundation. The rates per thousand are 100.89 and 47.08 for base-set and counter-set respectively.

The power of the Zeta test is seen once more in the fact that, at 66.43 per 1,000, the rate for Marvell's 'First Anniversary' surpasses that of every other specimen examined. Here, as with the Waller poems in Table 4, the test ranks the authentic text ahead of more than forty rivals. Of the main eight poems, moreover, the 'Fourth' and 'Fifth Advice', the two not considered to be Marvell's, rank below all of his. These are all much stronger results than might reasonably have been expected of a test treating of the middle frequency stratum. In that hitherto neglected stratum, as I remarked at the beginning of this article, the demands of subject and occasion might be expected to prevail over the effects of authorial habit.

For the Marvell satires, however, the central panel of Table 5 shows the effect of those very demands and is a stern reminder that the orientation of the Zeta test is not necessarily—and therefore not always—authorial. Apart from 'The First Anniversary', the poems associated with Marvell are all outscored by many texts by various other authors.

This sudden outcrop of failures needs to be understood.[8] Although Marvell's corpus is diverse in literary form, little of the unquestioned work that comprises his base-set is satirical. Several of his longer poems bear on affairs of state but his most persistent note is pastoral. Given the stipulation of consistent recurrence employed here, his word-list for the Zeta test includes many words that occur more often in pastoral than in other literary forms. Poems of that cast are marked not only by some loosely related lexical words but also by a tendency to use function words that were already becoming archaic. The twenty most common words of Marvell's Zeta-list are: flowers, doth, green, grass, lest, unto, who (interrogative), straight, O, Heaven's, pure, grief, Oh, thine, hence, roses, equal, stay, under, trees. Many of these twenty word-types (like others from further down the list) are more at home in Marvell's pastoral lyrics than in his satires where they occur, as a group, at less than half the rate they attain in his main set. Taken individually, several of them do show frequencies normal for Marvell while others do not occur in any of the political satires associated with him. The leading examples of the first sort are 'lest', 'straight', 'who', and 'under'. These can be taken as representing many other words, from further down the list, words that Marvell is always inclined to use. The leading examples of the second sort, which do not

occur in any of these satires, are 'grass', 'pure', and 'roses'. Other members of my little set of twenty, like 'green' and 'trees', put in an appearance or two. These, too, can be taken as representing a large number of non-political words from further down the list.

It must be emphasized that we are dealing not in absolutes but in lexical probabilities. Most English words can put in an appearance in quite unexpected contexts. But such instances are heavily outnumbered, in any given context, by words more usual there. Our language is so often metaphorical in cast that even a satire on naval affairs is not proof against the language of pastoral. Here is a fragment from my 'Dutch Comfort, or Our late Reverses at Sea'. Even here, the unexpected words soon begin to be outnumbered by those that might be expected. And, thinking of what is likely rather than what is conceivable—thinking, that is, statistically—one would expect more sails than trees, more guns than meadows in any long poem of this kind:

> Far o'er the Atlantick Meadows, pure and green,
> A threat'ning Clowd of Trees was to be seen.
> Close-haul'd, but favour'd by the mounting Gale,
> Van Tromp's main Squadron of some forty Sail!
> Our Duke, who strode the Poop with haughty
> Mien
> Had been outwitted by a Foeman keen.
> Proud Phoebus but a Phaëthon, we found,
> Our Fates with his inextricably bound.

Consider the four word-types that I purposely implanted here: pure, green, trees, and meadows, together with the singular forms of the two nouns. All told, they occur six times in the 14,198 words of the three 'Advice' poems associated with Marvell. In the other 'Advices', which comprise 2,312 words, they do not occur at all. They occur only sparsely in three of Marvell's long poems, 'A Poem upon the Death of O. C.', 'An Horatian Ode', and 'Fleckno, an English Priest at Rome'. Of their 4,503 words, these word-types make up only six. Three of them, all from 'On the Death of O. C.', are instances of 'tree(s)'. The rest are scattered single instances. But Marvell's 'Upon Appleton House' sets all these figures in high relief. In that poem, they occur twenty-seven times in 4,845 words.

In situations of this kind, little is gained by altering the stipulations so as to modify the word-list. To relax them is to weaken their power to differentiate. To tighten them is likely to intensify a given effect. Of the forty or so word-types that occur in all five segments of Marvell's main set, many are redolent of pastoral. It is fair to suppose that any pastoral poem, from *Lycidas* to Pope, would be likely to achieve a high score in this particular exercise and that no satire, from Dryden to Pope, would be likely to match it.

Ultimately, of course, the question is not whether I am justified in supposing that the dearth of 'pastoral' words in the 'Advices' is the main reason why the Zeta test fails to identify three of those poems as Marvell's. The question is what to make of a test where perfect accuracy is suddenly offset by utter failure. Our three failures make it clear that Zeta scores are not always reliable authorial markers. Our three successes, in each of which the scores for a specified poem surpass the scores for more than forty others, suggest that the Zeta test should not be discarded. That position is supported by the fact that other trials I have made have shown a high level of accuracy.

The difficulty, it seems, is that there is an intractable weakness in our 'model', a select word-list that does not reflect the full range of Marvell's repertoire. In such cases, a particular test may need to be set aside. It is as if a doctor were to say, 'I'm pretty sure this is a new mutation. If I'm right, there is no point in using our standard test this time. We'll just have to try another approach'. Far better for him to do so than to go on against the grain and discover the truth in an autopsy. Most poets usually write within their customary repertoires. Whenever they do so, the Zeta test shows high levels of accuracy. When they do not, a reader should know it. The language of the three main 'Advices' is Marvellian enough to enable the Delta and Iota tests to operate successfully. The content-words of the middle frequency stratum set these poems apart from the main body of his work. Now whereas even a good reader cannot easily see trends in the frequency of very common words, any reader of Marvell will recognize this large shift of subject.

But is it appropriate to appeal to the reader's judgment in this way? What of the argument that computational methods are meant to settle questions on which informed readers disagree? To make that a *sine qua non* is an unreasonable demand. Computational methods can often shed new light on vexed questions. But, like all other forms of inductive reasoning, they yield inferences, and not absolute proofs. The best inferences will be drawn by the best students of the full range of evidence: with new methods, as with old, we are always obliged to read and think. When disagreement persists, our best recourse is a return to the beginning. (One of Wittgenstein's dicta comes to mind: 'Back to the rough ground. Look and see.') It may be possible to find a weakness in reasoning, to form better inferences, or to show that the burden of proof is altered by the new evidence. It may be necessary to frame the question better and test it afresh. In twenty-five years, however, I have scarcely seen a case where well-founded computational work remained seriously at odds with the scholarly consensus.

5 Conclusions and Suggestions

One way of measuring the accuracy of these tests is to declare an error whenever the score for a text by the target-author is surpassed by that of any other test-piece. In Tables 4 and 5, the members of Group A, being components of the counter-set, should not be used for this purpose. Where they are involved, an 'error' is certainly a danger-signal: but it would be specious to register 'successes' of this kind.

For Waller, therefore, in the present case, we are measuring two poems of his authorship against forty-seven others. The results of the Zeta test yield no errors out of ninety-four comparisons. So, too, for the Iota test. At no point does either of these rather characteristic pieces encounter a near-rival. Matched against forty-five poems by other authors, Marvell's 'First Anniversary' also yields a perfect record of success for both Zeta and Iota. Its margin of advantage, however, is often much less than Waller's was. The Iota test registers forty-five successes out of forty-five for the 'Last Instructions' and the 'Second Advice', forty-three out of forty-five for the 'Third Advice'. (Even this last score becomes forty-five out of forty-five when the second stipulation is tightened.)

All told, the Iota test yields 272 successful comparisons out of 274, a success-rate of over 99%. The Zeta test is 100% successful for the two Waller poems and 'The First Anniversary'. With the political satires, however, the Zeta test is unable to distinguish Marvell from many of his fellows. The errors reach double figures for all three poems and the test is clearly inappropriate in such cases.

The one-on-one trials illustrated in Table 3 have yet to yield a genuine error. But, as the only aberrant score in that set indicated, such trials are better not undertaken until wider-ranging tests (or firm external evidence) have identified the main claimants. Stipulations designed to distinguish one author from one other cannot be expected to work properly on a third.

Further trials will show whether our only cluster of errors can be set aside as the effect of a predictable mismatch. If it can, we are looking overall at very high success-rates. That offers good reason for undertaking further work. And the Iota results, in particular, add to previous evidence that three, but only three, of the 'painter satires' are the work of Andrew Marvell.

Whoever undertakes them (myself I hope included), such further trials should obviously incorporate a range of variations on the stipulations employed here. Experience so far suggests some limits. It is always desirable to frame the second stipulation, which governs the counter-set, so as to exclude all the words that are used in common-words analyses. It is sometimes necessary to tighten this stipulation further in order to differentiate between authors. If the first stipulation, which governs the base-set, imposes too strict a rule of consistency few words will survive. When (as was noted earlier) the field is impoverished, the test is open to adventitious effects. Too strict a rule of consistency can also yield errors with genuine but uncharacteristic texts. Another worthwhile variation, for use with texts too short to allow the effective use of both Zeta and Iota, is what one might call 'Iota plus'. In this situation, the select list might exclude only those words that occur in, say, four or more segments of the base-set and also

exceed a fairly strict quantum for the counter-set. Like Iota itself, this approach yields a completely accurate differentiation between our four Marvell poems and all the rest. For short texts, then, it might well offer a useful check on the common-words procedures.

Apart from short texts and over-strict stipulations, aberrant frequencies most often arise in words that might reasonably be culled. Henry Fielding is more given than his sister to the adjective 'simple'. But her novel, *David Simple*, has a spectacular effect upon the word-count. The compelling objection to *ad hoc* forms of culling is that they make it too easy to tip the balance in one direction or another. In recent articles (e.g. 2004), David Hoover successfully culled those word-types that exceed a given frequency in any of the specimens. That approach seems an effective way of controlling aberrant behaviour in any of the specimen-texts without favouring a preferred outcome.

Further trials should also establish whether it is necessary, for the best effects, to work with text-sets of uniform length. The obvious advantage is offset by a sort of extravagance. This is less damaging with drama and prose fiction than with poetry and personal letters. Poetic texts and poetic corpora both vary too much in length to lie easily upon a procrustean bed. To obtain sets of uniform size, we must either reduce the larger to match the smaller or turn away from the smaller. Too little Milton and Congreve or no Rochester and Sedley? Put like that, the choice is too invidious to contemplate. A better way to put it is that we need to determine what is most appropriate, accepting only such limitations as we must, and resisting them when we can.

With so many variations still in play, it is too soon to offer a binding definition of a Zeta or an Iota score. It is better, at least for the present, to think of Zeta and Iota as working-labels for two little families of tests based upon select word-lists. The lists are formed on the basis of stipulated contrasts between a base-set and a counter-set. In Zeta, the stipulations admit only those word-types that attain a specified level of consistency in the base-set while failing to reach a specified level, whether of consistency or of frequency, in the counter-set. Iota rests upon the residue, embracing word-types that do not meet the first stipulation while occurring even more rarely in the counter-set.[9]

The results offered in this article, along with those obtained in other trials, suggest that these tests themselves may be of real use in cases of doubtful authorship. Although no two word-frequency tests can ever be entirely independent of each other, each of these tests employs frequencies from a different stratum and, again, from a stratum other than that used in the more familiar common-words procedures. For good and ill, however, they are very simple additive measures. Their ultimate value may be less direct, lying rather in demonstrating that evidence of authorship is indeed present in every frequency stratum. Our task now is to find the best ways of deploying it.

Acknowledgements

The author is indebted to those who reviewed this article for many helpful comments and suggestions. In their different ways, as on many similar occasions, Hugh Craig and Harold Love have given invaluable assistance and encouragement.

References

Burrows, J. (2005). Andrew Marvell and the ''Painter Satires'': A computational approach to their authorship. *Modern Language Review*, **100**: 281–97.

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

Burrows, J. (2002). Delta: a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**: 267–86.

Forsyth, R., Holmes, D., and Tse, E. (1999). Cicero, Sigonio, and Burrows: investigating the authenticity of the *Consolatio. Literary and Linguistic Computing*, **14**: 393.

Foster, D. W. (1989). *Elegy by W. S.: A Study in Attribution*. Newark: N.J.

Holmes, D. (1994). Authorship attribution. *Computers and the Humanities*, **28**: 96–8.

Hoover, D. (2004). Testing Burrows's ''Delta''. *Literary and Linguistic Computing*, **19**: 453–75.

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

Lord, G. deF. (ed.) (1963). *Poems on Affairs of State: Augustan Satirical Verse*, Vol. I, New Haven and London: Yale University Press, 1660–1678.

Margoliouth, H. M. (ed.) (1971). *The Poems and Letters of Andrew Marvell*, Vol. I, Oxford: Clarendon Press, 3rd edn.

McCarty, W. (2005). *Humanities Computing*. London: Palgrave.

Patterson, A. (2000). Lady state's first two sittings: Marvell's satiric canon. *Studies in English Literature*, **40**: 395–411.

Smith, N. (ed.) (2003). *The Poems of Andrew Marvell* London: Longman.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*, Cambridge MA: Addison-Wesley.

Notes

1 The provenance of the main 'Advice' poems is set out in Margoliouth (3rd ed., 1971), where only the 'Last Instructions' is accepted as Marvell's. The most accessible set of texts is in Lord (1963).

2 The corpus of about 540,000 words has been reduced by some six thousand by setting aside three poems by Waller and Marvell for use as independent specimens. The corpus 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 added by Christopher Wortham assisted by Joanna Thompson.

3 Like the greater part of the main database, some of the independent texts were entered, by keyboard, from standard texts. The rest were downloaded from the Chadwyck-Healey archive, to which my university subscribes. I am much in debt to all those whose work has made mine possible.

4 It is often useful to distinguish between word-types and word-tokens. The many occurrences of 'the' in any given text are called word-tokens, instances of the word-type 'the'.

5 See Burrows (2002, 2003) and Hoover (2004). A 1999 article (first shown to me in 1998) has a brief passage in which absolute z-scores are used as a measure of distance. Although I do not remember noticing it (in an article whose weight lies elsewhere), it may have given me a clue that came to mind when it was needed. I am content, in any case, to yield precedence on this point. See Forsyth, Holmes, and Tse (1999, 393).

6 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, 57–8).

7 My colleague, Hugh Craig, has been experimenting successfully on these lines.

8 Willard McCarty has been a persistent and eloquent advocate of the idea that we learn most from our failures. One must hope that it applies to the following remarks. See, for example, McCarty (2005), 286, s. v. 'failure'.

9 The notion of constructing an authorial word-list and then excluding those used by a stipulated number of other writers has not, I think, been tried. But there is an extensive literature on rare words, especially those that occur only once or twice in a given text. David Holmes (1994) offers a helpful summary of these procedures (among others). More recently, access to electronic archives has assisted in identifying words peculiar to a given author. The work of Donald W. Foster offers notable examples, embracing both success and failure. The protracted controversy surrounding a failed Shakespeare attribution (Foster, 1989) need not occupy us here.