Who wrote *Shamela*? Verifying the Authorship of a Parodic Text

John Burrows
University of Newcastle, Shortland, Australia

Abstract

Imitative texts of high quality are of some importance to students of attribution, especially those who use computational methods. The authorship of such texts is always likely to be difficult to demonstrate. In some cases, the identity of the author is a question of interest to literary scholars. Even when that is not so, students of attribution face a challenge. If we cannot distinguish between original and imitation in such cases, we must always concede that an imitator may have been at work. *Shamela* (1741) has always been regarded as a brilliant parody. When it is subjected to our standard common-words tests of authorship, it yields mixed results. A new procedure, in which special word-lists are established according to a predetermined set of rules, proves more effective. It needs, however, to be tried in other cases.

1 The Problem

The question, as posed above, is a mere flourish. *Shamela* (1741) was published anonymously, one of many parodies of the first volume of Samuel Richardson's *Pamela*, which had appeared in the previous year. Henry Fielding never acknowledged it as his, and his authorship has sometimes been disputed. But it was an open secret in his own days and the modern scholarly consensus leaves no genuine room for doubt. Behind the flourish, nevertheless, lies a genuine question. Are our computational tests of authorship able to identify the author of an imitative text? The matter embraces serious imitations, where the intent is to deceive, as well as parodies, where the intent is to satirize. Texts revised or extended by a second author stand at the edge of the same territory. In all these cases, the identification of the real author can be of intrinsic scholarly interest. Even when, as with *Shamela*, that is not so, the question of our capacity remains. Unless we can identify the true authors of such works, every identification we offer is open to the objection that an imitator may have been at work. *Shamela* makes an appropriate point of entry into this whole area because it has always been recognized for its parodic brilliance. It is no surprise to find that it does, in fact, offer a stern challenge to several current tests of

Correspondence:
John Burrows, CLLC,
University of Newcastle,
Shortland, NSW 2308,
Australia.
E-mail:
john.burrows@netcentral.com.au

Literary and Linguistic Computing, Vol. 20, No. 4, 2005.
doi:10.1093/llc/fqi049

authorship. But a rather different kind of test, with which I have lately been experimenting, assigns it unambiguously to Henry Fielding.

2 The Texts

In order to test these matters more thoroughly, I have turned away from the large set of 'histories' on which much of my computational work on prose fiction has been based for many years. These retrospective narratives, couched in the first person, provided an unusually homogeneous set of specimens, relatively free of the sharp stylistic contrasts that obtain between third-person narrative and dialogue. But, as a compensation for the loss of this stylistic homogeneity, the new set of texts to be used here ranges further across the works of Fielding and Richardson, and their more immediate contemporaries. In keeping with the recent practice, which makes it easier for others to replicate one's procedures, the new set of texts is substantially unmodified. Two sorts of alteration should be mentioned—ancillary matter, like chapter titles and page numbers, was enclosed in angle brackets to exclude it from the word-counts and single quotation marks were converted to doubles in order to distinguish them from apostrophes.

Table 1 lists the new datasets for Fielding and Richardson. In each case, there is a main body of about a hundred thousand words for use as a procedural basis and a further range of independent test specimens. (A similar division obtains for the texts of other authors, which will be set out later in this article.)[1] The main sets are designed to represent each author's prose fiction. They also include a single specimen apiece of their non-fiction. Fieldimg's *Essay on Conversation* is a discursive piece. Richardson's *Familiar Letters* is a set of model letters for use in many social situations such as the dismissal of an employee or the rejection of a proposal of marriage.

Of the test specimens, Fielding's range widely across his prose works including some essays, and his personal letters. (These last are distinct from his *Familiar Letters*, which, unlike Richardson's, are from a work of prose fiction. They are five letters contributed by him to his sister Sarah's *Familiar Letters between the Characters in 'David Simple'*.) The Richardson test specimens are chosen on a different basis. In order to focus on comparing *Shamela* with Fielding's particular target, they comprise successive 20,000-word segments from the first volume of *Pamela*. Whether or not *Shamela* can be seen as more broadly Richardsonian is a further point of interest.

3 The Evidence of Very Common Words

This broader question is an appropriate place to begin because it can easily be accommodated in another process, that of ensuring that the two main sets can be distinguished on authorial lines. Each of the main sets was split, for this initial purpose, into ten segments

1 Except where otherwise indicated, the new texts were downloaded from the Chadwyck-Healey archive of electronic texts, to which my university subscribes. The texts were used only to extract word-counts.

Table 1 List of Fielding and Richardson texts

MAIN SETS	
Henry Fielding	
27412	*Joseph Andrews,* Book iv (Chadwyck-Healey)
15110	*Jonathan Wild,* Book iii (*ibid.*, hereafter C-H)
27439	*Tom Jones,* Book i-ii (C-H)
18720	*Amelia,* Book viii (C-H)
11978	*Essay on Conversation* (Gutenberg)
100659	
Samuel Richardson	
24021	*Pamela,* last part of Vol. i (C-H) cf. Everyman edn., pp. 406–53
27417	*Clarissa,* Letters 1–13 (C-H)
22666	*Grandison,* Letters 1–12 (C-H)
26025	*Familiar Letters,* Letters 61–135 (C-H)
100129	
TEST SPECIMENS	
14144	*Shamela* (C-H)
Samuel Richardson	
197035	*Pamela,* bulk of Vol. i (Gutenberg) cf. Everyman edn., pp. 1–405 (treated as ten segments of 20,000 words)
Henry Fielding	
24954	*Joseph Andrews,* Book i (C-H)
27720	*Joseph Andrews,* Book ii (C-H)
24228	*Joseph Andrews,* Book iii (C-H)
15659	*Jonathan Wild,* Book i (C-H)
15473	*Jonathan Wild,* Book ii (C-H)
14338	*Jonathan Wild,* Book iv (C-H)
22972	*Tom Jones,* Book xii (C-H)
19286	*Tom Jones,* Book xiii (C-H)
18598	*Amelia,* Book x (C-H)
16972	*Amelia,* Book xii (C-H)
10821	*Familiar Letters (*S. Fieldimg), xl-xliv (Garland)* (H. Fielding's acknowledged contribution)
6628	*Female Husband* (U. of Virginia etexts)
13066	*Journey from this World* (Garland)*
3526	*The Opposition* (Garland)*
16674	*Covent Garden Journal,* Nos. 60–62, 64–72 ed. G. Jensen (NY, Russell, 1964).*
17105	*Letters of Henry Fielding* (per fav. Clive Probyn)

*text entered by keyboard from specified source.

of 10,000 words. In the event, the main distinction offered no difficulty. But, as to *Shamela*, the results of these first comparisons were mixed.

The outcome of several cluster analyses seemed mutually contradictory. In every trial, there was a sharp and accurate separation between the entries for Fielding and those for Richardson. But, in different trials, *Shamela* shifted from one camp to the other. The variation, it emerged, was governed by the length of the word-list. The shorter the list, the more *Shamela* resembled Richardson. With longer lists, it moved into Fielding's territory. This pattern deserved to

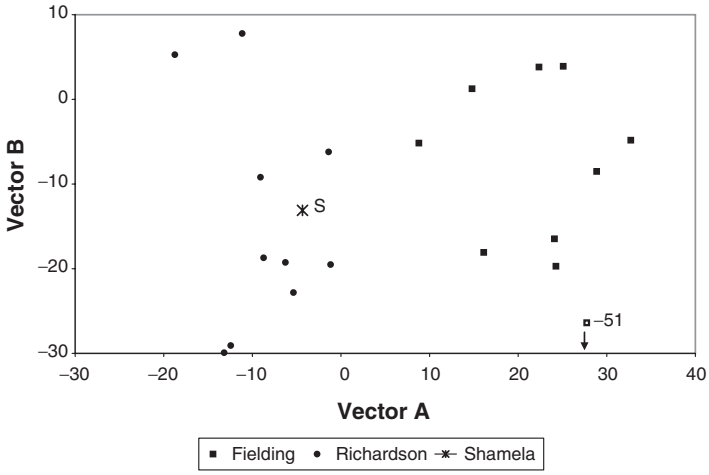

Fig. 1 Principal component analysis based on the 75 most common words of the main Fielding

be kept in mind because it might reflect the increasing proportion of lexical words in longer word-lists. And that, in turn, might imply that Fielding captured more of Richardson's syntax and major rhetorical habits than of his vocabulary at large.

Principal component analysis has the advantage of cluster analysis in admitting more than a single perspective. On its horizontal axis, Fig. 1 shows a clear separation between the clusters of Richardson and Fielding entries, with *Shamela* lying towards the nearer edge of the former cluster. Its nearest Fielding neighbour is an entry for *Amelia*. On the vertical axis, the topmost entries are both for *Pamela*, followed by two entries for *Joseph Andrews*. At the bottom lie two entries for Richardson's *Familiar Letters* while the entry for Fielding's *Essay on Conversation* lies (as the arrow indicates) far beyond the frame. On this axis, *Shamela* lies near the centre of all the entries, with entries for *Clarissa* just above it and entries for *Sir Charles Grandison* below it.

A scatter-plot for the word-variables themselves (as represented in the first two principal components) underlies Fig. 1. It gives added meaning, therefore, to that picture of the text specimens. It is not reproduced here because it is easy to describe. The location of the personal pronouns and the common inflected verbs around the outskirts of the plot has most influence upon its shape. *I*, *my*, and *me*, accompanied by *you* and *your*, lie close together at the far left, offset at the far right by *the* and *which*. The third-person singular pronouns and the common past-tense verbs lie at the top (with the titles *Mr* and *Mrs* not far below them). At the bottom are *we, our*, *it*, *an*, and *this*, together with the present-tense verb forms. The major prepositions range across the 'South-Eastern' corner of the plot. Other common words enrich this pattern but most of them lie further from the outskirts and exert less influence on the shape of the whole group.

To traverse the vertical axis of Fig. 1 from top to bottom, it follows, is to move from texts in which reported narrative predominates to more discursive forms of writing. ('He did' and 'she said' stand

opposed to 'we are' and 'it may be'.) On this basis, of course, the entries for works of non-fiction should lie lowest of all, as indeed they do. The much more powerful horizontal axis, meanwhile, separates the authentic texts in accordance with their authorship; and it puts *Shamela* with Richardson. Fielding's imitation, on this analysis, seems a complete success.

But such a straightforward authorial reading of Fig. 1 is facile. Richardson's fiction is always epistolary. It is couched, accordingly, in the first person with frequent recourse to the particular *you* to whom a given letter is addressed. Fielding, by contrast, unites impersonal narration with dialogue: his fiction, accordingly, covers the gamut of the common verbs and pronouns. There are sufficient grounds here for a marked separation between the two authorial clusters. But, on this evidence, one can scarcely say more of *Shamela* than that, being epistolary like *Pamela*, it is bound to show a *prima facie* resemblance to Richardson. Fig. 1, in short, is not false but it is inadequate.

To employ the Delta procedure, where a specimen is matched against the work of many authors, we must turn away, for the moment, from the direct contest between Fielding and Richardson. Several trials of this kind with *Shamela* were a little more helpful. Neither Richardson nor Fielding was ever far from the head of the field of candidates, nor seized the lead. An appropriate outcome for a hybrid text, one might suppose—but still no clear indication of its authorship. Not quite like Fielding because his imitative attempt is successful enough to obscure his own signature. Not quite like Richardson because the attempt is not a complete success. But the Delta procedure requires that one of its many candidates must lead the field. The fact that Daniel Defoe emerged as the most consistent leader was suggestive. Of all the 18th century male novelists tested, Defoe and Richardson are most akin in their divergence from the formal stylistic patterns of their classically educated fellows. Their vernacular, *ad hoc* styles of first-person reportage differ from each other—but much less than they differ from the rest. It is here that Fielding finds his target and, in doing so, moves far enough from his usual stylistic repertoire to defeat the tests employed so far:

> Mrs. *Jervis* and I are just in Bed, and the Door unlocked; if my Master should come—Odsbobs!! I hear him just coming in at the Door. You see I write in the present Tense, as Parson *Williams* says. Well, he is in Bed between us, we both shamming a Sleep, he steals his Hand into my Bosom, which I, as if in my Sleep, press close to me . . . (Henry Fielding, *Shamela*, Letter VI)

4 The Evidence of Less Common Words

4.1 Fielding vs. Richardson; Richardson vs. Fielding

In recent years, common-word procedures like those used above have played a leading part in computational tests of authorship and have

yielded many valuable results. The back-files of *Literary and Linguistic Computing* and *Computers and the Humanities* offer abundant examples. To use these methods, however, especially with short texts, is to discard so much of one's material that it is difficult to escape a sense of thriftlessness.

Meanwhile, at the opposite end of the frequency-spectrum, there is a centuries-old history of tests treating of words that can reasonably be regarded as the property of a given target-author. If they also occur in a text of doubtful authorship, they can be adduced as evidence. The logic is impeccable but its practical implementation is beset with difficulties. If the grounds on which words are accepted as the target-author's property are too weak, the evidence is of little weight. But, if the grounds are strong, so few words pass muster that they cannot be expected to occur, in convincing numbers, in the doubtful text. We all have our pet expressions; but we do not use all or even most of them whenever we pick up a pen. Their absence from a given text is therefore of less evidential weight than their presence. And even their presence does not count for much except through a cumulative impact. The heart of the difficulty is that words are not often peculiar to any one writer. Even a fresh coinage, like Shakespeare's 'incarnadine' does not long remain his property. When it does recur, it is not in any work of his.

But let us envisage a different approach, forsaking the quest for words *peculiar* to a given writer's work and looking, instead, for those that recur there more often than they do in other writers. A little rule might be designed to collect all the words that recurred to a stipulated extent in a sizeable 'base set' of the target-author's work. An accompanying rule would then exclude those among them that occurred too freely in the 'counter-set' of work by an appropriate group of other authors. A simple variant of the procedure (with which we shall begin) could be employed in cases where only two authors were involved. The object, one might say, is to give up the search for nuggets of pure gold and try processing lower-grade ore.

The 100,659 word-tokens of my main set of Henry Fielding's writings embrace 7977 word-types. If this base set is broken into five successive segments of 20,000 word-tokens, 2,247 of the word-types occur in three or more segments. Of these, 1,109 word-types occur no more than five times, all told, in the 100,129 word-tokens of Richardson's counter-set. At this point, a little cautious culling of the list was undertaken in order to remove word-types whose frequency patterns were aberrant enough to disturb the outcome. They included proper names and words used principally as titles rather than as common nouns, *colonel, serjeant, doctor*, and *squire* among them. I culled *thee* and *thy* because they, too, occur often enough to create special difficulties.

Finally and rather uncertainly, I culled *hath* but not *doth*. These two words have long been recognized as characteristic of Henry Fielding, used freely by him at a time when they were already archaic.

Hath occurs 188 times in his main set, not at all in Richardson's main set, and twice in our 197,035-word selection from *Pamela*. The corresponding figures for *doth* are fifty-two, one, and one. *Hath* occurs fifty times in the 14,144 words of *Shamela* and *doth* fifteen. It seemed to me that, while Fielding's use of *hath* in *Shamela* is a sign of his authorship, its high frequency there would give it undue weight among the occurrence rates we shall be assessing. (With *hath* omitted, *book*, at a frequency of twenty, was the highest single score remaining for *Shamela*). All told, fifteen word-types, were culled, leaving an initial working-list of 1094. In hindsight, much of this culling made little difference. But cases where the matter is of some weight can arise unexpectedly. *Simple* is a word favoured by Henry Fielding. When Sarah Fielding uses it as the name of a protagonist, its frequency pattern runs amok. Instead of making questionable decisions, case by case, about words like *hath* and *simple*, it may be best to adopt David Hoover's more objective scheme of culling all word-types that occur disproportionately often in any one test-specimen.[2]

Raw frequencies for the chosen word-list are extracted from each of the test specimens and set beside those for the base set and the counter-set in a single work-sheet. (Microsoft Excel is well-suited to the purpose.) In the present case, the first work-sheet embraces 1094 rows for the chosen word-types and twenty-nine columns of actual data for the specimens listed in Table 1. Fielding supplies the base set of 100,659 word-tokens, Richardson the counter-set of 100,129. There is a single entry for *Shamela*, followed by ten for the segments of *Pamela*, and sixteen for the Fielding specimens. The behaviour of the base set and the counter-set is heavily constrained by their role in establishing the word-list. But the other twenty-seven specimens are free agents.

The total of all the raw frequencies in each column is calculated and then turned into a rate per thousand of the word-tokens in each text. After that, it is easy to sort, cut, and recalculate so as to vary the stipulations. The stipulation for the base set can be tightened to include only those word-types that occur in four or even all five of the Fielding segments. The counter-set can be tightened to exclude those word-types that occur five or four or three times, twice, or even once in Richardson. Lax restrictions do not yield an effective contrast between the authors under scrutiny. Unduly tight restrictions reduce the word-list to a point where the numbers are too fragile to carry much evidential weight.

Table 2 summarizes the outcome of a series of tests in which the threshold for the base set was held steady at three segments out of five while the number of occurrences allowed in the counter-set was progressively reduced from five to zero. This had the effect, as shown, of reducing the word-list, step by step, from 1,094 to 289. The constrained contrast between base set and counter-set, registered as rates per thousand words, began as 80.94 against 20.66. In the final step, it stood at 18.33 against zero.

2 See, for example, his 'Testing Burrows's Delta', *Literary and Linguistic Computing*, 19 (2004), 456–7.

Table 2 Fielding, Richardson, and *Shamela*. Frequency rates for sets of uncommon words

Stipulations: 1. Base set: 3 ex 5 segments. 2. Counter-set: N ex 100129									
Stipulated maxima in counter-set				**0**	**1**	**2**	**3**	**4**	**5**
Number of compliant word-types				**289**	**528**	**717**	**874**	**993**	**1094**
				Rates of occurrence (tokens per 1000)					
Base set: Fielding			100659	18.33	35.84	49.03	60.35	71.44	80.94
Counter-set: Richardson			100129	0.00	2.39	6.16	10.87	15.62	20.66
TEST SPECIMENS									
Richardson	Pamela1		20000	1.70	5.05	9.45	13.40	18.35	23.50
	Pamela2		20000	1.40	4.85	9.50	13.90	19.95	25.20
	Pamela3		20000	3.40	7.20	12.00	16.35	23.20	29.45
	Pamela4		20000	2.95	7.50	14.65	20.35	27.75	33.90
	Pamela5		20000	3.15	6.30	10.80	15.60	23.30	28.50
	Pamela6		20000	3.45	6.90	11.55	16.00	20.80	26.55
	Pamela7		20000	1.80	5.35	9.20	13.85	18.15	24.65
	Pamela8		20000	2.25	5.30	9.10	13.60	17.85	23.60
	Pamela9		20000	3.05	6.70	10.30	13.90	18.00	23.60
	Pamela10		17035	2.35	5.17	9.22	13.50	19.20	25.89
		Max		**3.45**	**7.50**	**14.65**	**20.35**	**27.75**	**33.90**
		Mean		**2.55**	**6.03**	**10.58**	**15.05**	**20.65**	**26.48**
		stdev		**0.75**	**0.99**	**1.77**	**2.17**	**3.21**	**3.31**
Fielding	JABk1		24954	11.74	25.25	34.42	43.56	52.26	60.63
	JABk2		27720	12.88	25.36	36.22	44.99	52.24	61.40
	JABk3		24228	13.74	26.42	38.01	46.85	56.05	65.50
	JWBk1		15659	12.71	30.08	41.57	50.83	58.94	67.25
	JWBk2		15473	14.41	30.83	43.75	52.09	60.49	69.86
	JWBk4		14338	13.11	31.73	45.75	54.19	62.42	71.84
	TJBk12		22972	13.32	25.81	36.31	45.62	54.72	62.51
	TJBk13		19286	10.99	23.18	33.34	42.83	51.44	59.63
	AmelBk10		18598	12.96	24.20	31.19	38.39	46.13	52.21
	AmelBk12		16972	12.14	24.10	33.05	42.01	50.61	56.45
	FamLett		10821	11.46	23.10	33.36	41.59	51.01	57.85
	FemHusb		6628	11.62	23.08	32.44	39.98	49.49	57.48
	Journey		13066	13.32	27.17	38.27	46.76	55.79	66.28
	Opposn		3526	12.20	26.09	34.88	47.08	56.44	67.50
	Covent		16674	11.51	22.97	34.48	43.48	55.12	63.63
	Letters		17105	8.48	20.29	29.41	37.77	45.07	50.04
		Min		**8.48**	**20.29**	**29.41**	**37.77**	**45.07**	**50.04**
		Mean		**12.29**	**25.60**	**36.03**	**44.88**	**53.64**	**61.88**
		stdev		**1.38**	**3.12**	**4.50**	**4.69**	**4.80**	**6.14**
Shamela			**14144**	**5.94**	**15.55**	**22.70**	**30.90**	**38.11**	**43.91**
z-score (vs. *Pamela*)				**4.53**	**9.57**	**6.86**	**7.31**	**5.43**	**5.26**
z-score (vs. Fielding)				**−4.59**	**−3.22**	**−2.96**	**−2.98**	**−3.24**	**−2.93**

As for the test specimens, the outcome is clear. In every column except the first from the left (where the stipulations imposed may simply be too strict), *Shamela* leans towards Fielding, in a pattern favouring his authorship. A close scrutiny of the individual scores, a comparison of averages, and a weighing of z-scores all point that way. The very low standard deviations for the test specimens show that each set of data is cohesive. And the range of z-scores implies that the outcome is of considerable statistical weight. Those for the contrast

between *Shamela* and Fielding, lying around –3.00, show a clear divergence from his usual vocabulary. Those for the contrast between *Shamela* and *Pamela*, ranging up from +5.00, register sharply different populations. So far as these tests are concerned, therefore, *Shamela* lies at the edge of Fielding's lexical repertoire but quite beyond Richardson's. Student's *t*-test and discriminant analysis might be used to verify these comments, but it seems unnecessary to labour the obvious.

Let us turn instead to the obverse case, taking Richardson as base set and Fielding as counter-set. Richardson's 100,129 word-tokens embrace 7063 word-types. When the set is split into five segments, 1,856 of these word-types occur in three or more of the five. Exactly 700 of them occur five times or fewer in Fielding's counter-set. After six words were culled, the remaining 694 were treated in the same fashion as before.

As Table 3 shows, the inversion of the process reinforces the original result. With Richardson as base set, the scores for *Pamela* are now in the ascendant. Those for Fielding lie far below them. Those for *Shamela* lie in between, but are always closer to Fielding. Once again, the standard deviations are so low (in proportion to their various mean scores) that the data can be seen as highly cohesive. And, if the distance between *Shamela* and the mean scores for each author's set of test specimens is measured in z-scores, the previous proposition is surpassed. *Shamela* now lies within Fielding's range but beyond Richardson's. A comparison between the z-scores in Table 2 and those in Table 3 yields a further point. As a parodist, not unexpectedly, Fielding is more successful in capturing elements of Richardson's vocabulary than in abandoning his own.

4.2 Fielding vs. nineteen others

In the face of the evidence offered so far, a scholar—even one who believed that Samuel Richardson might laugh at himself—would seem ill-advised to argue that *Shamela* is Richardson's. But, while this evidence distinguishes parody from target, it does not adequately address the question of Fielding's authorship. We have yet to test the claim that it is not his work. Although the idea need not be pursued too strenuously, it is at least worth illustrating how a real problem of this sort might be approached.

The immediate requirements are, first, an adequate main set of texts by other novelists who were at work around the middle of the 18th century and, second, an adequate set of test specimens by them and also by some others. As in the previous case, my old set of 'histories' have been forsaken and fresh texts are used for the purpose. They are listed in Table 4.

Within the limits of availability, the main set of texts in Table 4 was chosen to range across the prose fiction of the mid 18th century at a rate of about 20,000 words per author. The set of test specimens was more carefully designed. Except for Bage's *Hermsprong*, these texts

Table 3 Richardson, Fielding, and ***Shamela***. Frequency rates for sets of uncommon words

Stipulations: 1. Base set: 3 ex 5 segments. 2. Counter-set: N ex 100659									
Stipulated maxima in counter-set				**0**	**1**	**2**	**3**	**4**	**5**
Number of compliant word-types				**155**	**288**	**395**	**508**	**613**	**694**
				Rates of occurrence (tokens per 1000)					
Base set: Richardson			100129	10.00	22.49	28.99	37.53	46.00	52.76
Counter-set: Fielding			100659	0.00	1.32	3.45	6.82	10.99	15.01
TEST SPECIMENS									
Richardson	Pamela1		20000	3.15	11.30	17.25	22.20	28.75	35.50
	Pamela2		20000	3.65	10.80	16.80	22.30	26.55	34.10
	Pamela3		20000	4.15	9.95	15.40	21.25	25.65	33.45
	Pamela4		20000	4.25	12.55	17.15	23.40	30.20	38.60
	Pamela5		20000	3.30	9.65	14.45	21.30	27.40	34.90
	Pamela6		20000	4.85	12.15	18.75	25.75	31.05	38.75
	Pamela7		20000	5.05	10.40	17.00	23.40	29.70	36.50
	Pamela8		20000	5.25	14.05	22.05	29.30	35.70	43.00
	Pamela9		20000	5.35	12.75	18.25	24.30	30.40	38.85
	Pamela10		17035	3.52	11.39	17.79	23.36	29.29	36.69
		Min		**3.15**	**9.65**	**14.45**	**21.25**	**25.65**	**33.45**
		Mean		**4.25**	**11.50**	**17.49**	**23.66**	**29.47**	**37.03**
		stdev		**0.83**	**1.38**	**2.04**	**2.41**	**2.80**	**2.85**
Fielding	JABk1		24954	2.12	4.61	7.69	12.06	16.87	22.28
	JABk2		27720	1.30	3.50	7.11	10.39	15.66	20.71
	JABk3		24228	1.11	4.25	7.51	10.94	16.34	20.84
	JWBk1		15659	1.72	4.41	7.79	12.90	17.82	23.69
	JWBk2		15473	1.29	4.14	7.30	11.18	14.67	19.19
	JWBk4		14338	1.53	3.56	6.70	11.44	16.39	22.04
	TJBk12		22972	1.65	3.96	6.44	9.84	14.71	19.37
	TJBk13		19286	3.21	5.65	9.23	13.07	17.58	22.30
	AmelBk10		18598	1.83	4.19	6.56	10.81	15.06	18.71
	AmelBk12		16972	1.47	3.54	6.36	9.60	12.67	16.50
	FamLett		10821	2.50	7.49	10.07	13.31	20.05	25.41
	FemHusb		6628	1.06	4.83	7.24	12.07	17.95	20.82
	Journey		13066	1.45	3.37	6.20	10.41	14.85	19.75
	Opposn		3526	0.85	2.55	6.81	9.93	17.58	21.84
	Covent		16674	2.40	6.42	10.14	13.91	18.77	25.97
	Letters		17105	1.99	5.90	9.53	13.62	17.89	22.57
		Max		**3.21**	**7.49**	**10.14**	**13.91**	**20.05**	**25.97**
		Mean		**1.72**	**4.52**	**7.67**	**11.59**	**16.55**	**21.37**
		stdev		**0.61**	**1.28**	**1.33**	**1.43**	**1.88**	**2.45**
Shamela			**14144**	**1.70**	**5.16**	**10.11**	**14.35**	**20.01**	**25.45**
z-score (vs. *Pamela*)				**−3.07**	**−4.59**	**−3.61**	**−3.86**	**−3.38**	**−4.07**
z-score (vs. Fielding)				**−0.04**	**0.50**	**1.83**	**1.93**	**1.84**	**1.67**

also fell within about 30 years on either side of *Shamela*. (Bage was added because he, like Smollett, is thought to resemble Fielding.) Nine of the fourteen authors are represented in the main set, six by selections from works other than those used in the main set and Richardson by selections from each of his novels. In order to see whether entirely independent agents would perform differently, the group also includes five whose work does not figure in the main set. (These are Amory, Bage, Paltock, Swift, and Walpole.) The full set

Table 4 Main set of texts by nineteen authors, with sixteen further test specimens

MAIN SET		
22680	1739	Aubin, Penelope, *Lucinda*, pp. 163–243
19797	1719	Barker, Jane, *Exilius*, pp. 1–89
20913	1763	Brooke, Frances Moore, *Lady Julia Mandeville*, pp. 1–163
23341	1765–70	Brooke, Henry, *The Fool of Quality*, pp. 1–146
21030	1749	Cleland, John, *Memoirs of a Woman of Pleasure*, pp. 3–124
23355	1727	Davys, Mary, *The Accomplish'd Rake*, pp. 1–88
21846	1724	Defoe, Daniel, *Roxana*, pp. 1–67
19804	1747	Fielding Sarah, *Familiar Letters between the Characters* . . ., xxxiii–xxxix
21792	1766	Goldsmith, Oliver, *The Vicar of Wakefield*, ch. i–xv
20192	1773	Graves, Richard, *The Spiritual Quixote*, pp. 196–305
20726	1761	Hawkesworth, John, *Almoran and Hamet*, ch. i–xiii
21353	1725	Haywood, Eliza, *The Injur'd Husband*, pp. 121–80
18695	1759	Johnson, Samuel, *Rasselas*, Vol. ii, ch. xxvi–xlviii
20522	1752	Lennox, Charlotte, *The Female Quixote*, Book ii
20013	1720	Manley, Mary, *The Power of Love*, Novels ii and iii
19425	1741	Richardson, Samuel, *Familiar Letters*, Letters 1–60
21131	1767	Sheridan, Frances, *The History of Nourjahad*, pp. 1–186
20478	1748	Smollett, Tobias, *Roderick Random*, ch. xi–xvii
20727	1768	Sterne, Laurence, *A Sentimental Journey*, Vol. Ii
397820		
TEST SPECIMENS		
9386	1756	Amory, Thomas, *The Life of John Buncle Esq.*, pp. 1–43
9351	1796	Bage, Robert, *Hermsprong*, Vol. i, ch. i–v. (Blackmask Online)
7319	1722	Defoe, Daniel, *Moll Flanders*, pp. 1–22
18955	1759	Fielding, Sarah, *The History of the Countess of Dellwyn*, Vol. iv
8658	1773	Graves, Richard, *The Spiritual Quixote*, pp. 306–52
6670	1725	Haywood, Eliza, *The Injur'd Husband*, pp. 202–220
8148	1759	Johnson, Samuel, *Rasselas*, Vol. i, ch. i–x
14430	1751	Lennox, Charlotte, *The Life of Harriot Stuart*, Vol. ii, pp. 1–63
11529	1714	Manley, Mary, *The Adventures of Rivella*, pp. 1–60
7414	1751	Paltock, Robert, *Peter Wilkins*, ch. i–iv
24021	1740–41	Richadrson, Samuel, *Pamela,* cf. Table 1, *supra*
27417	1747–48	Richardson, Samuel, Clarissa, cf. Table 1, *supra*
22666	1753–54	Richardson, Samuel, *Sir Charles Grandison*, cf. Table 1, *supra*
7701	1751	Smollett, Tobias, *Peregrine Pickle*, ch. i–iv
7684	1726	Swift, Jonathan, *Gulliver's Travels*, Vols. i–ii
7843	1764	Walpole, Horace, *The Castle of Otranto*, ch. i

The text of Sarah Fielding's *Familiar Letters* was entered by keyobard from microfilm. The texts used for Richardson's *Pamela* and Bage's *Hermsprong* were downloaded from the Gutenberg Project and Blackmask Online respectively. The remainder were downloaded from the Chadwyck-Healey Archive of 18th Century Fiction. The extent of my debt is obvious, and I am happy to acknowledge it.

comprises sixteen specimens to balance the original set of test specimens by Fielding.

For the trials whose outcome is set out in Table 5, Fielding supplies the base set of 100,659 word-tokens. Since the first stipulation (that a word-type occur in at least three of the five 20,000-word segments) is unchanged, we begin as before with 2,247 word-types. The counter-set now embraces 397,820 word-tokens by nineteen authors. In order to accentuate authorial individuality, the second stipulation was modified

Table 5 Fielding, nineteen authors, and *Shamela*. Frequency rates for sets of uncommon words

Stipulations: 1. Base set: 3 ex 5 segments. 2. Counter-set: N ex 19 authors								
Max. contributors from counter-set				**1**	**3**	**5**	**7**	**9**
Number of compliant word-types				**61**	**191**	**381**	**604**	**838**
				Rates of occurrence (tokens per 1000)				
Base set: Fielding			100659	3.45	10.43	21.16	**36.33**	55.18
Counter-set: 19 authors			397820	0.15	1.48	5.34	12.66	23.03
TEST SPECIMENS								
Fielding	JABk1		24954	1.40	5.57	14.11	25.45	40.59
	JABk2		27720	1.05	4.40	12.37	24.42	38.60
	JABk3		24228	1.49	6.60	13.74	25.09	39.25
	JWBk1		15659	2.55	6.45	12.90	24.97	42.47
	JWBk2		15473	1.94	5.95	12.34	26.82	43.69
	JWBk4		14338	2.37	6.63	13.11	26.36	46.17
	TJBk12		22972	1.04	5.92	12.28	22.59	36.31
	TJBk13		19286	0.78	4.20	10.16	21.36	33.65
	AmelBk10		18598	0.81	3.39	7.42	20.97	33.93
	AmelBk12		16972	1.24	3.77	9.31	22.51	35.88
	FamLett		10821	1.39	7.21	14.97	25.69	40.20
	FemHusb		6628	0.91	3.92	9.51	17.65	30.63
	Journey		13066	2.22	5.59	12.09	24.34	40.49
	Opposn		3526	1.13	3.69	12.48	23.26	33.47
	Covent		16674	1.13	4.81	12.42	22.94	38.87
	Letters		17105	1.34	4.38	10.06	19.88	31.75
		Min		**0.78**	**3.39**	**7.42**	**17.65**	**30.63**
		Mean		**1.42**	**5.15**	**11.83**	**23.39**	**37.87**
		stdev		**0.56**	**1.23**	**2.00**	**2.51**	**4.44**
Other authors	Amory		9386	0.85	2.45	9.06	17.05	30.15
	Bage		9351	0.32	3.42	8.34	16.79	29.73
	Defoe		7319	0.82	1.37	5.19	9.15	16.67
	S Fielding		18955	0.74	2.69	6.65	15.30	28.33
	Graves		8658	0.69	2.54	6.81	15.71	29.68
	Haywood		6670	0.30	1.95	6.30	11.54	18.89
	Johnson		8148	0.74	2.70	5.89	15.71	27.61
	Lennox		14430	0.07	1.80	6.31	13.44	23.70
	Manley		11529	0.61	2.95	6.25	13.01	23.77
	Paltock		7414	0.40	2.83	7.42	14.30	24.55
	Richardson1		24021	0.17	1.04	4.58	10.66	20.44
	Richardson2		27417	0.33	2.08	6.67	14.01	24.22
	Richardson3		22666	0.44	2.60	5.87	11.38	21.44
	Smollett		7701	1.04	3.12	7.92	17.14	30.91
	Swift		7684	0.65	1.56	6.90	15.88	27.59
	Walpole		7843	0.38	1.79	5.48	14.41	25.76
		Max		**1.04**	**3.42**	**9.06**	**17.14**	**30.91**
		Mean		**0.53**	**2.31**	**6.60**	**14.09**	**25.21**
		stdev		**0.27**	**0.67**	**1.16**	**2.41**	**4.28**
Shamela			**14144**	**1.98**	**4.45**	**10.75**	**19.30**	**29.98**
z-score (vs. Fielding)				**1.00**	**−0.57**	**−0.54**	**−1.63**	**−1.78**
z-score (vs. others)				**5.32**	**3.19**	**3.56**	**2.16**	**1.11**

so as to exclude word-types not used by more than a given number of the nineteen.

After *hath* was culled, the second stipulation was applied so as to exclude word-types used by more than nine of the nineteen. This

yielded 838 word-types as the word-list for a first trial, whose outcome is recorded in the right-hand column of Table 5. As before, the strong contrast between the scores for main set and counter-set (55.18 vs. 23.03 per thousand) means no more than that the behaviour of those two sets is constrained by the chosen stipulations. As before, the scores for the two sets of test specimens are successfully distinguished. The mean scores for the two sets stand well apart: 37.97 for Fielding, 25.76 for the counter-set. Standard deviations of 4.44 and 4.28 show that the difference between the mean scores is is of real weight. At 30.91, Smollett's score, the highest of the sixteen, is the only one to exceed the lowest of the sixteen Fielding entries, 30.63 for *The Female Husband*, a sort of 'documentary fiction'.

So far, so good. But, at 29.98 per thousand, the score for *Shamela* falls neatly between the mean scores for the other test specimens. The z-scores show that, while it is closer to the sixteen than to Fielding, it is within the ordinary range of divergence for each of them. The result is as inconclusive as those yielded by the common-word tests employed at the beginning of this article.

As the second stipulation is progressively tightened, it becomes clear that the original allowance was too lax. In the second column from the right of Table 5, bearing on the 604 word-types that occur in no more than seven of the nineteen authors, *Shamela* lies a little closer to Fielding than to the rest and even the highest of them falls short of *Shamela* and of any Fielding entry. The z-scores, however, are still too weak to yield a firm conclusion.

The trend strengthens in the next two columns, treating of word-types that occur in no more than five and three, respectively, of the nineteen authors. The mean scores stand in sharp contrast. The ultimate diversity of the counter-set is beginning to manifest itself in relatively high standard deviations and some of the higher scores surpass Fielding's minimum while falling short of *Shamela*. The z-scores no longer show any meaningful divergence from Fielding but the divergence from the others is very marked. The scores shown in the remaining column of Table 5 treat of the sixty-one word-types that occur in no more than one of the nineteen authors. They extravagantly favour Fielding's authorship. But by now we are beginning to deal in frequencies too low for this approach. In the last column, even so, we see further evidence that Henry Fielding was indeed the author of *Shamela*. And, by the same token, we see a promising new way of identifying the authors of imitative texts whose provenance is less secure.

Figure 2 highlights these results and shows their strength. It can be left, by now, to speak for itself. But it should not be regarded as a substitute for the more detailed evidence of Table 5. It shows, more vividly than the Table, that *Shamela* resembles Fielding more closely than it resembles the generality of his fellows. The Table adds, for example, the crucial point that it resembles Fielding more closely than it resembles *any* of the others.

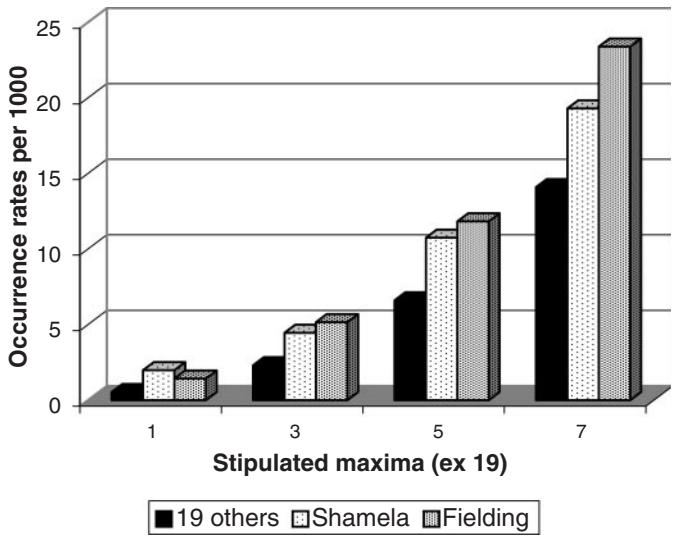

Fig. 2 Scores of *Shamela*, Fielding, and nineteen others

With only sixty-one word-types still qualifying, the left-hand column of Table 5 is an appropriate point to abandon our present approach. Beyond this, it is better to forsake overall word-counts and examine the words themselves. At the extreme, only two word-types occur in all five segments of the base set and nowhere in the nineteen-author counter-set. They are *doth* and *lastly*, which occur fifty-two and fifteen times respectively in the base set. They occur fifteen times and twice, respectively, in *Shamela* and, like *hath*, help to mark the authorship of that work.

As mentioned earlier, *hath* and *doth* are recognized as words used freely by Henry Fielding when most others had forsaken them. It is possible, it seems, to put *lastly* beside them, though rather as a word whose day was yet to come. After being detected as a word that does not occur in all the 397,820 word-tokens of our counter-set, it is no wonder that it is a rare bird among the sixteen test specimens. Of those 199,192 word-tokens, it yields only two—one in Sarah Fielding and one in Swift. A potent enough indicator to offer us a little *coda*? Not altogether. While its presence is indeed a sign of Henry Fielding, it is not a reliable indicator: it occurs in only ten of the sixteen Fielding test specimens. Its effect is actually more potent when it is allied with others, mostly less striking in themselves. When they are gathered in sufficient numbers according to appropriate stipulations, they yield the compelling results shown in the central columns of Table 5. And that is no mere flourish.