Testing Burrows's Delta

David L. Hoover
New York University, USA

Abstract

Delta, a simple measure of the difference between two texts, has been proposed by John F. Burrows as a tool in authorship attribution problems, particularly in large 'open' problems in which conventional methods of attribution are not able to limit the claimants effectively. This paper tests Delta's effectiveness and accuracy, and shows that it works nearly as well on prose as it does on poetry. It also shows that much larger numbers of frequent words are even more accurate than the 150 that Burrows tested. Automated methods that allow for tests on large numbers of differently selected words show that removing personal pronouns and words for which a single text supplies most of the occurrences greatly increases the accuracy of Delta tests. Further tests suggest that large changes in Delta and Delta z-scores from the likeliest to the second likeliest author typically characterize correct attributions, that differences in point of view among the texts are more significant than differences in nationality, and that combining several texts for each author in the primary set reduces the effect of intra-author variability. Although Delta occasionally produces errors in attribution with characteristics that would normally lead to a great deal of confidence, the results presented here confirm its usefulness in the preliminary stages of authorship attribution problems.

1 Introduction[1]

First in his Busa Award presentation, and then in two recent articles, John F. Burrows has presented a simple new measure of stylistic difference that seems very promising for studies of authorship attribution, especially those in which the range of possible claimants cannot easily be narrowed by traditional methods (2001, 2002a, 2003). He has also applied the new measure to the stylistics of English translations of Juvenal (2002b). The new unitary measure of authorial difference, which Burrows calls 'Delta,' is based, like many other measures and techniques, on differences in the frequencies of the most frequent words in a group of texts. In his initial discussions of Delta, the texts analyzed are selections of 'verse by twenty-five poets of the English Restoration period' (2003, p. 10). Burrows uses the frequencies of the 150 most frequent words of the entire set of texts in his exposition of the method. After the frequencies of all these words in all of the texts are collected, he calculates the mean frequency for each word in the entire set and compares it with the word's frequency in a test

Correspondence:
David L. Hoover
Department of English
New York University
19 University Place, 5th Floor
New York, NY 10003, USA.

E-mail:
david.hoover@nyu.edu

[1] I would like to thank an anonymous reviewer for suggestions that improved the clarity and force of my argument.

text and in the selection by one of the authors in the primary set. The result is two differences that are then compared with each other.

Delta is a relatively simple measure of difference, but its calculation and interpretation are not very transparent. In the interest of clarity, it seems worthwhile to trace through an example. Consider the word *of*, which is the third most frequent word in the texts that Burrows analyzes, with a mean frequency of 1.821 (presumably, this is its percentage of all the tokens) (2002a, p. 272). For *Paradise Lost*, its frequency is 2.769, and for the selection by Behn from the main set, its frequency is 1.783. Delta compares how different the two texts are from the mean of the corpus, and here those differences are .948 for *Paradise Lost* and –.038 for Behn's selection, showing that Milton uses *of* much more frequently and Behn slightly less frequently than the mean for the corpus. Given how quickly word frequencies drop from the most frequent words to the hapax legomena, it is important to convert these absolute differences in frequency to z-scores (by subtracting the mean frequency of the word in the corpus from its frequency in the test text and dividing this difference by the standard deviation of the word in the corpus). This transforms the raw word-frequency information into a measure of the distance (in standard deviations) of each frequency from the mean frequency for the corpus, and shows that *Paradise Lost* is 3.015 standard deviations above the mean, and Behn's selection is .121 standard deviations below the mean. When Behn's z-score is subtracted from the z-score for *Paradise Lost* to determine the difference between the differences from the mean, the result is 3.136, showing that *of* is used very differently in *Paradise Lost* than it is in the works of Behn. When this procedure is followed for all 150 words, the result is a list of differences between the differences of the two texts from the mean, and the stage is set for the final step in the calculation of Delta. Because Burrows is interested in the pure differences between the differences, he eliminates their signs before calculating their mean. This result is Delta: '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' (2002a, p. 271). After Delta is calculated for each pair of texts, the primary author who shows the smallest mean difference from the test text, the smallest Delta, is the likeliest author of the text.

In spite of the simplicity of Delta and the fact that its calculation systematically removes information about the direction of the differences, it is remarkably effective in identifying authors in a difficult 'open' test. When texts by 16 authors who are members of the original set and texts by 16 other authors are tested with Delta to determine likely authorship, 'Of thirty-two long poems, . . . fifteen are correctly identified and another 15 yield scores that correctly place them outside the main set' (Burrows, 2003, p. 15). Although these results are not completely accurate, they are very encouraging, suggesting that Delta can be a powerful tool in the early stages of authorship studies in which there are many possible claimants. Because of Delta's great potential, further tests seem appropriate—tests that shift the focus to prose and investigate how the accuracy of Delta in attributing texts to their correct authors and in eliminating authors in the main set as claimants is affected by different methods of selecting and limiting the word-frequency lists, by nationality, by point of view, and by different numbers of texts.

2 The Design of the Investigation

The primary set of texts for my first test on prose consists of sections of pure authorial narration taken from the beginnings of twenty third-person American novels published between 1890 and 1925.[2] This choice of dates ensures a large selection of texts for testing: many such novels, now out of copyright, are readily available as e-texts.[3] After cleaning up the e-texts to correct for problems of hyphenation, apostrophes and single quotation marks, and any other peculiarities, I created samples of approximately 25,000 words of pure authorial narration from each text by manually deleting the dialogue and very short narrative passages (roughly, fewer than 30 words). Selecting only third-person narration for analysis eliminates the effects of different proportions of dialogue and narration among the novels and prevents the (often considerable) differences among the voices of the characters from clouding the issue of authorship.

The resulting prose sections range from just over 10,000 words to just under 39,000 words, the shortest sections coming from short novels with a great deal of dialogue, in which cases they include almost all of the narration of the novels. Both the median and mean size of the twenty sections is approximately 27,000 words, and the entire corpus consists of 539,089 words—approximately the same size as the corpus of poetry analyzed by Burrows, though representing only 20, rather than 25 authors.[4] The secondary set of texts, created as above, consists of 25 sections of third-person authorial narration by members of the first set of authors and 14 by other authors, and contains a total of 1,027,319 words.[5]

A combination of text-analysis tools and custom programming was used to create several word-frequency lists for the primary set of texts using various criteria of selection that will be described below.[6] Burrows's analysis (2002) shows that the accuracy of Delta decreases when the frequency list is reduced from the 150 to the 40 most frequent words, and Hoover (2001, 2002, 2003) shows that cluster analyses based on as many as the 800 most frequent words are usually more accurate than those based on the smaller lists that have traditionally been tested; therefore, the 800 most frequent words were collected for analysis here.

Delta was calculated for each text by a Microsoft Excel spreadsheet that accepts as input sets of columns of the most frequent words from up to 80 primary and 80 secondary texts. A macro selects the most frequent words of each primary and secondary text in turn and calculates and records Delta for various numbers of them, beginning with the 800 most frequent words, and continuing with the 700, 600, 500, 400, 300, 200, 150, 100, 70, 50, 30, and 20 most frequent. Once all of the sets of words

[2] The texts in the primary set are listed in the appendix.

[3] As an anonymous reviewer for *Literary and Linguistic Computing* has pointed out, the texts analyzed in this article are nearly 100 years old. The fact that Burrows's texts date from approximately 200 years earlier strongly suggests that the date of the text is not a crucial factor, but tests on contemporary texts, if successful, would provide additional strong support for the general usefulness of Delta as a measure of authorship. I have recently begun such tests on a body of contemporary American poetry of approximately the same size as Burrows's Restoration sample (with 25 primary authors, 15 texts by members, and 15 by others). Preliminary results show that Delta is even more effective on contemporary American poetry, correctly attributing all 15 of the texts by primary authors and correctly suggesting that the other 15 texts were written by authors who are not in the primary set. Furthermore, it does so consistently across a very wide range of analyses. More investigation is needed, but the success of Delta on poetry and prose written over a period of 300 years is very encouraging.

[4] All word-counts presented here must be considered approximate. Although intuitively *word* seems a simple concept, there is no one 'correct' number of words in a text of any significant length. The number of words depends on a series of decisions about what counts as

have been entered and Delta calculated and recorded for the 20–800 most frequent words for each of them, another macro determines and records the rank of the actual author (for texts by authors in the main set) and extracts and arranges the information about the accuracy of the identifications in a format appropriate for graphing (the spreadsheet, with macros, is available upon request).

This automation assures the consistency and accuracy of the analysis, saves countless hours of tedious and painstaking drudgery, and allows for the rapid testing of many differently selected sets of frequent words. For example, I tested the initial group of 39 secondary texts using the following eight different kinds of sets for each of the 13 different numbers of words mentioned above, for a total of 104 tests for each text and 4056 tests in all:

1. The most frequent words
2. Contractions removed[7]
3. Personal pronouns removed
4. Contractions and personal pronouns removed
5. Culled at 70%; that is, words for which a single text supplies more than 70% of the occurrences are removed; see Hoover (2002, p. 170), for discussion.
6. Culled at 70%; contractions removed
7. Culled at 70%; personal pronouns removed
8. Culled at 70%; contractions and personal pronouns removed

3 The Effectiveness of Delta for Pure Third-Person Narration in Fifty-Nine American Novels

An analysis based on the 20–800 most frequent words of all kinds for the first sets of texts shows that, as Burrows also found, the accuracy of Delta increases as the number of frequent words included in the analysis rises from the forty most frequent to the 150 most frequent (see Fig. 1). It also shows that the accuracy continues to increase at least up to the 600 most frequent words. When all eight of the different sets of words described above are analyzed (with the number of different results limited to make the graph easier to read), Delta correctly identifies the actual author of 23 of the 25 texts by members of the primary set, a result achieved five times (see Fig. 2). The best results are based upon far larger numbers of words than Burrows tested: for these texts, the maximum accuracy usually seems to occur with the 700 most frequent words. Removing contractions, personal pronouns, and culling at 70% produces two results in which 23 of the 25 attributions are correct, compared to at most one such result for any other set, as Fig. 2 also shows. The largest total number of correct attributions is achieved, however, by removing only personal pronouns and culling at 70%. The total numbers of correct attributions for each of the eight kinds of word sets, shown in Fig. 3, suggests that removing contractions generally reduces the accuracy of an

a word. Different text-analysis programs and different analysts produce different counts. There is no harm in this, as long as words are counted in a consistent way across all of the texts in any one analysis.

[5] The 25 texts by authors in the primary set are listed in the appendix.

[6] Unlike Burrows, I have not disambiguated any words in the texts analyzed here. (For example, Burrows separates the infinitive marker from the preposition in occurrences of the word *to*.) The very large amount of text involved in the tests and the large number of words to be analyzed made disambiguation impractical. Smaller groups of texts that had been disambiguated for other projects (Hoover, 2001, 2002) were tested, however. When a set of 10,000-word sections of pure third-person narration from novels was tested both normal and disambiguated, the disambiguated texts performed slightly worse than the normal ones. With smaller, 4,000-word sections of criticism, disambiguated texts performed somewhat better overall, though they produced the same maximum number of correct attributions as the plain texts did. Although disambiguation seems theoretically superior, these results suggest that Delta tests can reasonably be performed on plain texts.

[7] Not many contractions are frequent enough in pure narrative to be relevant. For the texts analyzed here, for example, only the following contractions appear among

analysis overall. Finally, note that all of the results in which there are 23 correct attributions involve culling at 70%, confirming the importance of assuring that words that are extremely frequent in only one of the texts—typically proper names—do not exert an undue influence on the analysis.

An examination of the incorrect attributions shows that, whenever Delta attributes 23 of 25 texts to their correct authors, it succeeds and fails for the same authors and the same texts. The failures are for Henry James's *The Europeans* and Ellen Glasgow's *Virginia*. Anyone familiar with the well-known differences in James's early and late styles will not be surprised that *The Europeans* of 1878 is not very similar to *The Ambassadors* of 1909, nor, perhaps, that in these five tests Delta invariably suggests Edith Wharton as the most likely author of *The Europeans*. Wharton's relatively high Delta (approximately .79 in these analyses) reminds us that Delta simply tells us which author in the primary set is least unlike the test author, and, as Burrows remarks, '"least unlike" need not be very like' (2003, p. 15). The spreadsheet also calculates the z-score of Delta for each author. For Wharton, the z-scores for Delta are approximately –1.5 in the most accurate analyses. This means that Wharton's Delta is 1.5 standard deviations below the mean, and, therefore, that fewer than 7% of the results in a normal distribution would be smaller. That is, fewer than 7% of authors would be less different from the test author by

the 1,200 most frequent words: *can't, couldn't, didn't, don't, hadn't, wasn't,* and *wouldn't*. For tests without contractions, these words and the words that comprise them have been deleted from the word-frequency lists. Best practice would be to replace the contractions with the uncontracted forms, or to adjust the frequencies of all affected words so as to account for them. The small number of words affected, however, suggests that the simpler method employed here is acceptable.

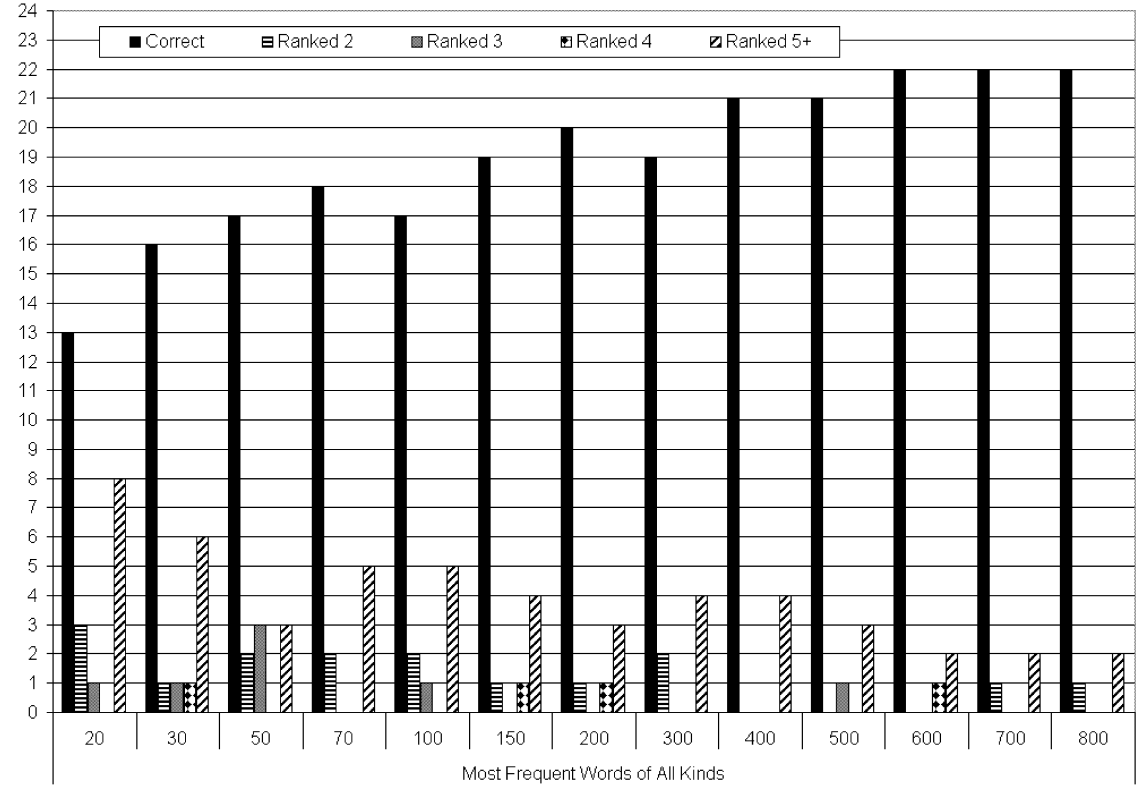

Fig. 1 Delta test results for the 20–800 words of all kinds in 59 American novels.

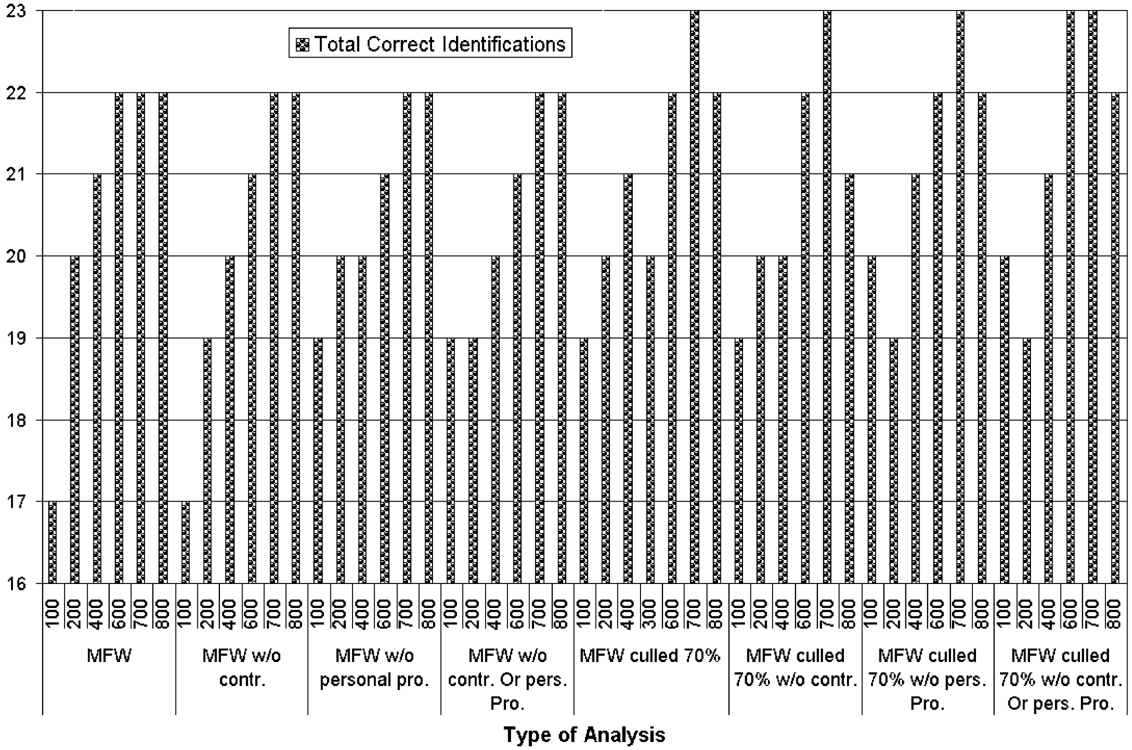

Fig. 2 Delta test results for eight sets of words in 59 American novels.

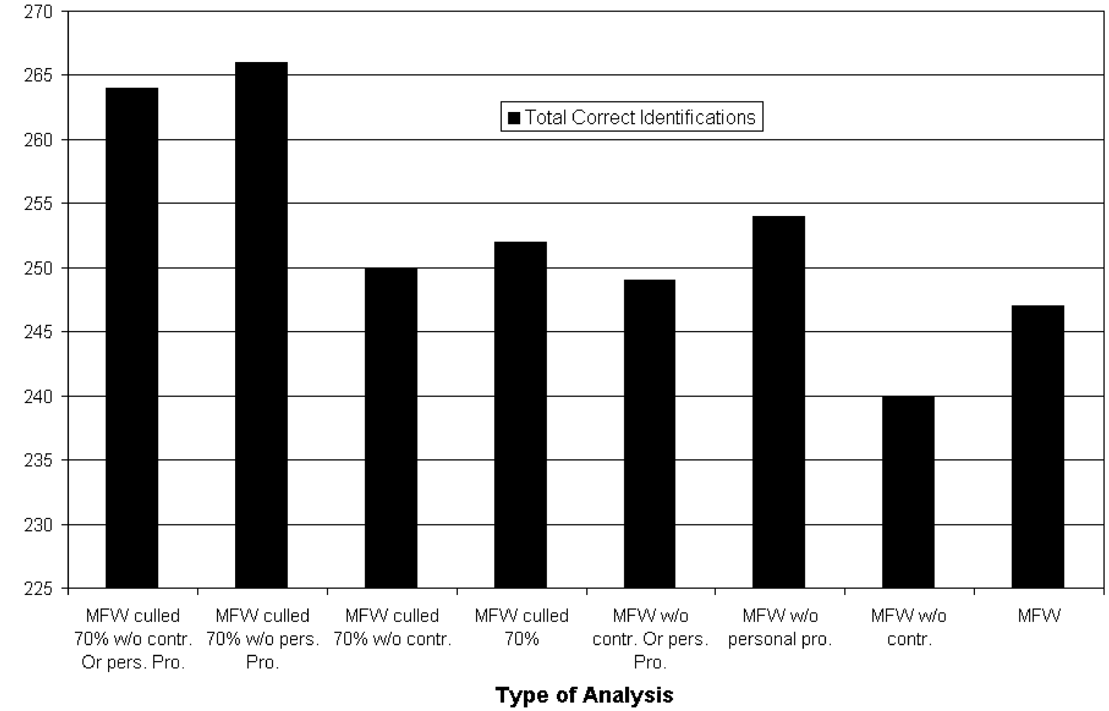

Fig. 3 Total correct attributions in Delta tests on eight sets of words in 59 American novels.

chance. (I will refer to the z-score of Delta as 'Delta-Z' below, and ignore its sign, referring to a Delta-Z of –3 as 'larger' than a Delta-Z of 1 and 'smaller' than a Delta-Z of –4.)

Although *The Europeans* was included intentionally to see whether the change in James's style over time would register as a difference in authorship, it would also have been legitimate, in my view, to have selected only early or late James because of this known anomaly.[8] Unfortunately, one can never know all such possible anomalies in advance, a fact that underscores the need for preliminary tests that can uncover problems before any authorship testing takes place. Note that James's late novel, *The Wings of the Dove* (1909), is always correctly attributed to him in the most accurate analyses. In fact, Delta successfully attributes *The Wings of the Dove* to James in 102 of 104 analyses (with Delta-Z's of approximately –2.6), and he is ranked second in the other two.

Ellen Glasgow's work is not as well known as James's, but *Virginia* (1913) 'was the first Glasgow novel considered by the author to be in her mature style' (Richards, 1971, p. 128). Perhaps, then, a contrast between it and the early *The Battle-Ground* (1902) should not be too surprising. Furthermore, two of Glasgow's other early novels, *Voice of the People* (1900) and *The Deliverance* (1904), are correctly attributed to her in all 104 analyses, strongly suggesting that 'early Glasgow', at least, is a bona fide style. In the five most accurate analyses, the author who ranks first as the likely author of *Virginia* is Edith Wharton (with Delta z-scores of approximately –1.6). The fact that Glasgow ranks only as high as eighth (once) as the likely author of *Virginia*, and as low as nineteenth (once), however, is cause for concern, suggesting, at the least, that the style of *Virginia* is statistically very different from the style of the earlier novels. Both of these cases of the failure of Delta to attribute texts to their correct authors should repay further study (already underway), and both suggest that Delta may be useful as a discovery technique for wide variations in style within an author's works, as well as in authorship attribution.

Examining the results of the analyses in which Delta correctly attributes 22 of 25 texts is also revealing. The only two texts besides *The Europeans* and *Virginia* for which Delta fails in these 16 analyses are Theodore Dreiser's *The Titan* and David Graham Phillips's *Susan Lenox*: seven failures and nine successes for Dreiser; nine failures and seven successes for Phillips. Significant differences might be expected between Dreiser's partially expurgated first novel from the primary set, *Sister Carrie*, which tells the story of a kept woman who eventually succeeds as an actress, and *The Titan*, part of a trilogy about the rise to power of a ruthless industrialist.[9] Furthermore, among these 16 analyses, Dreiser ranks no lower than third as the likely author of *The Titan*. In all seven failures, David Graham Phillips, a prolific muck-raker with concerns that are often congruent with Dreiser's, ranks first as the likely author. (His novel in the primary set, *The Conflict*, tells the story of an abortive romance between a labor-organizer and a member of the local aristocracy.) It is interesting that none of the nine analyses that succeed for

[8] In the difficult case of a text of unknown authorship that might be by early or late James, we might reasonably begin by including distinctive early and late texts and treating James as two possible authors. If our test text turned out to be least unlike early James or late James, we could then turn to more precise and powerful methods of attribution for further research.

[9] It is worth noting, however, that it is much easier to convince oneself that an incorrect attribution is reasonable once it has occurred than it is to predict which identifications will fail and which incorrect authors are least unlike the correct authors before the fact.

Dreiser and fail for Phillips are culled at 70%. This may seem surprising because the culling process typically improves the accuracy of analyses, but consider also that all five of the analyses with 23 correct attributions involve culled lists. Overall, Delta correctly identifies Dreiser as the most likely author of *The Titan* in only 38 of the 104 analyses and ranks him second in 33 more, third in 19, and fourth or higher in 14.

In all nine cases in which Delta fails to identify Phillips as the author of *Susan Lenox*, Tarkington is ranked as the most likely author. As in the case of Glasgow, Delta does an excellent job in attributing the other two early novels by Phillips, *The Fashionable Adventures of Joshua Craig* (1909) and *The Grain of Dust* (1911), correctly to him, with a total of 197 correct attributions in 208 analyses. The 11 times Delta fails, Phillips is ranked second as the most likely author 10 times and third once. Seven of these occur when only the 20 or 30 most frequent words are tested, and all analyses in which 400 or more words are tested succeed for both novels. (It should also be noted that *Susan Lenox* is generally considered Phillips's finest novel, and that it was published posthumously.)

For Restoration poetry, Burrows shows that Delta is usually lower for texts by authors from the main set than for texts by authors from outside the main set, as one would expect from a measure of difference. In his investigations, Delta ranges from .745 to 1.375 for texts by members of the primary set, and from 1.118 to 1.547 for the others, showing considerable overlap. There is almost no overlap in Delta-Z, however, which is larger than –1.9 for all but one of the texts by members of the primary set, and smaller than -1.9 for all but one of the texts by other authors (2003, pp. 15–20). Thus Delta is not only very effective in attributing texts to their correct authors, it is also very effective in eliminating the primary authors as claimants for texts by other authors.

The results for the American narrative fiction analyzed here are somewhat weaker: Delta is generally lower for texts by authors from the main set, ranging from .539 to .924, than for texts by other authors, ranging from .692 to 1.045, but there is a great deal of overlap. Delta-Z for texts by authors from the primary set ranges from –3.335 to –1.288 and for texts by other authors ranges from –2.319 to –1.357. Here there is more overlap than Burrows found, as can be seen in Fig. 4, which presents the results in much the same form as Burrows's Fig. 2 (2003, p. 20).

In addition to the two attribution errors for texts by authors in the primary set, Delta-Z is below the threshold score of –1.68 for two other texts by members of the primary set, and above the threshold for three texts by other authors. Furthermore, three of the texts by other authors have a Delta-Z greater than –2. As Burrows points out, such high z-scores must be taken seriously: they show that, at least for this set of authors and texts, false attributions are a serious possibility. In a normal distribution, only about 3% of texts would display Delta z-scores greater than –2, so that one would be justified in accepting such a z-score as strong evidence for authorship. As usual, Burrows's careful qualifications are salutary; he reminds us that 'the system for distinguishing between insiders and outsiders is not foolproof. It behooves us, as always, to remember that, by

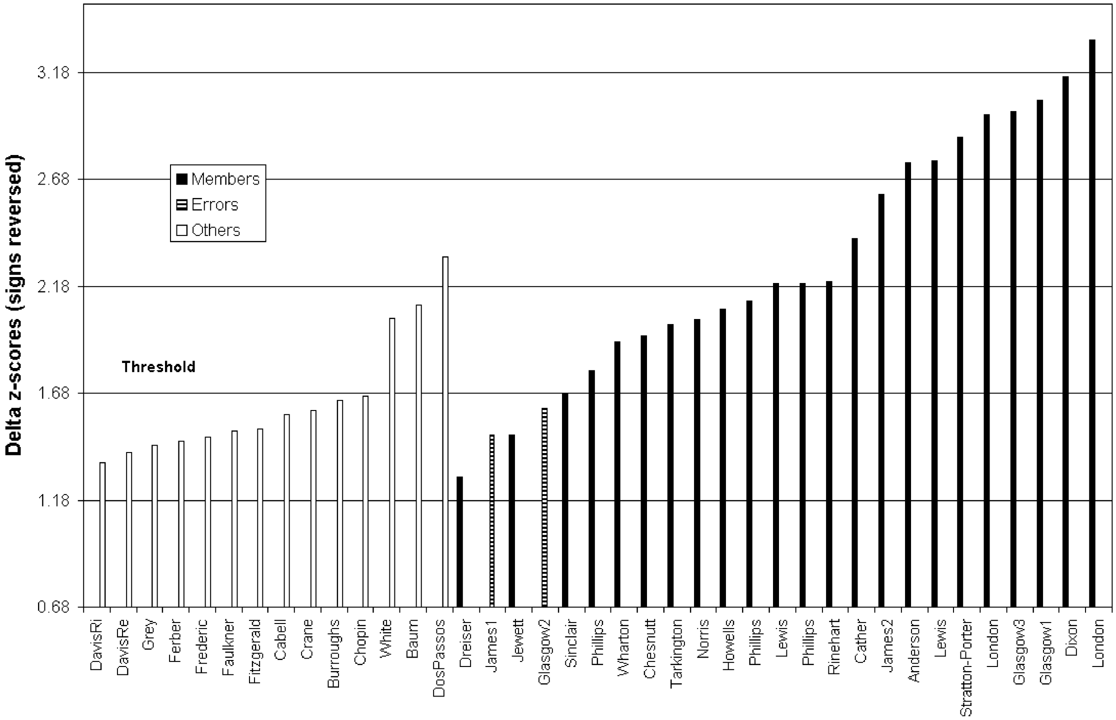

Fig. 4 Delta-Z for members, errors, and others.

relying on statistical analysis, even in this simple form, we are dealing in probabilities and not in absolutes.' (2002, p. 281).

The effectiveness of Delta in this kind of open-ended test can be further evaluated by comparing the results of a cluster analysis on the same texts. The results of analyzing all 59 of the texts in the primary and secondary sets above by cluster analysis (selecting the most frequent words of the entire corpus and then removing the personal pronouns and contractions and culling the frequency list at 70%) show that Delta gives better results than cluster analysis, which, in this case, incorrectly attributes 12 of the 59 texts. It fails to group Dreiser's two novels, incorrectly clusters Fitzgerald and Frederic, fails to cluster Glasgow's *Virginia*, James's *The Europeans*, and Phillips's *Susan Lenox* with these authors' other novels, fails to cluster either Jewett's or Tarkington's two novels, and fails to separate Twain's novel from others.

Finally, as a reminder of how much depends upon the initial choice of primary and secondary texts, consider what happens if the same 59 texts are analyzed again, but with different choices for primary and secondary texts. For this analysis, 18 of the same authors as before appear in the primary set (each represented by a different novel than before), Robins and Twain are moved to the secondary set, and Baum and Burroughs to the primary set. If the analyses that are the most successful with the initial set are repeated, Delta successfully attributes only 16 of the 25 texts by members of the primary set. The primary novels for this test are intentionally chosen so as to produce poor results, but one might often be faced with an analysis in which there is no known basis upon which to choose the primary and secondary texts, and nothing prevents an unfortunate set of texts like this from occurring by chance.

4 Delta and Delta-Z for Authors Ranked First and Second: Another Measure of Delta's Effectiveness

Burrows found that the likeliest author indicated by Delta tests for texts by authors not in the main set tended to change as the numbers of words included in the test was reduced from 150 to 60, while the likeliest author for texts by members of the main set tended to remain consistent. In a total of 80 tests for each of the two groups of 16 texts, the likeliest author changed for 14 of the 16 texts by other authors, but for only two of the 16 texts by members of the main set (2003, p. 23). The results for the prose analyzed here are again weaker overall, but the pattern of stability versus instability in attribution that Burrows observed is remarkable. When ten different numbers of words (70, 100, 150, 200, 300, 400, 500, 600, 700, 800) were tested for the 14 texts by other authors (140 total tests), 13 texts changed likeliest author. In contrast, in 250 tests on the 25 texts by members of the main set, only six changed likeliest author, and in 219 of these tests the true author ranked first or second. Furthermore, Glasgow's *Virginia* and James's *The Europeans,* which have been discussed above, account for 20 of the 31 tests in which the true author was not ranked first or second.

In the course of the testing above, one other difference between the results of Delta tests on the members of the primary set and those on other authors became clear: both Delta and Delta-Z for texts by members tend to drop much more rapidly for the second most likely author than they do for texts by others. This effect is quite visible in graphs of Delta or Delta-Z for the authors ranked first and second, but it is even easier to see in Fig. 5, which is based on the ratio of Delta-Z for the authors ranked first to Delta-Z for the authors ranked second. The strong tendency of Delta and Delta-Z to change more rapidly for texts by members than for texts by others is an additional indication that Delta is capturing genuine authorial characteristics.

Secondary texts by members tend to be very similar to their primary texts, and the real authors tend to be very likely, while the second most likely authors tend to be much less likely. The link between a secondary text by another author and the putative author's primary text, on the other hand, is relatively tenuous, and the authors ranked second and third are almost as likely as the one ranked first. Thus, if a claimant in an authorship test shows a low Delta and a high Delta-Z (relative to the scores of other authors in the test), if the likeliest author is consistent over a large range of numbers of frequent words, and if Delta and Delta-Z for the second-ranked author are very different than for the first-ranked author, the case for authorship may be considered strong.

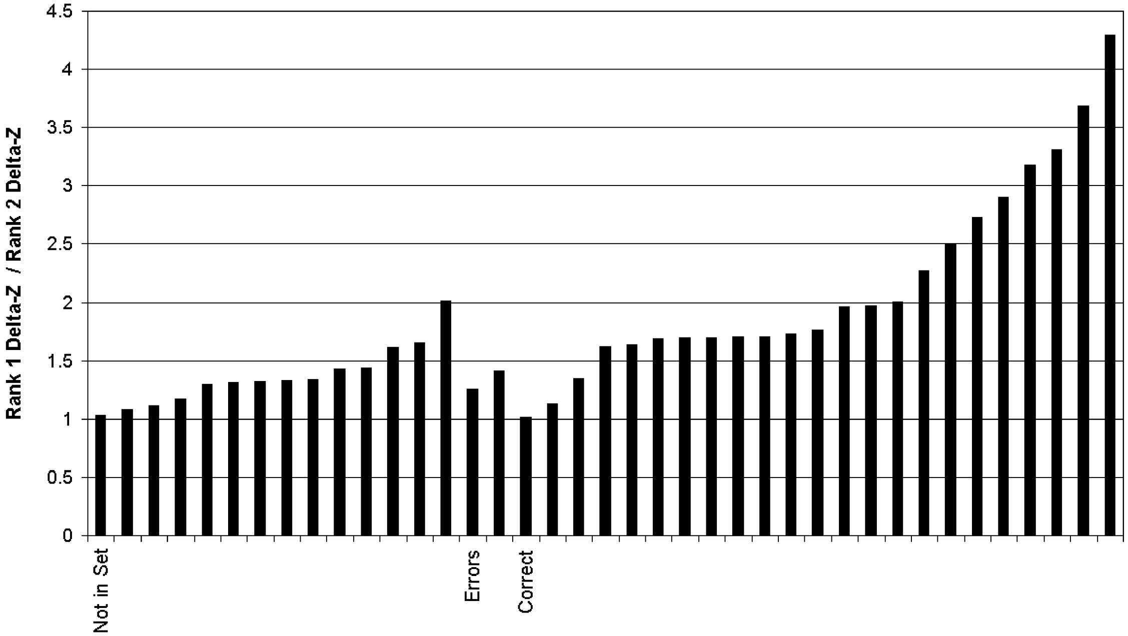

Fig. 5 Members, errors, others: rank 1 Delta-Z/rank 2 Delta-Z.

5 Further Evaluation of the Results of Delta Tests: Authorship Simulations

One way to evaluate the effectiveness of Delta in authorship attribution further is to assume for a moment that the authors of the second set of texts described above are not known, and to sort all of the texts into one series in increasing Delta-Z order, as shown in Fig. 6. Without any indication of which attributions are correct, it is not easy to decide where the threshold should fall. From Delta-Z alone, the attributions of the last eleven texts seem compelling, and, except for Dos Passos, they are correct. Table 1 gives the Delta, Delta-Z, and the percentage change in these two measures from rank 1 to rank 2 for a selection of the texts in Fig. 6. Note that for London and Dixon, at the far right in Fig. 6, not only is Delta-Z very high, Delta is low, and both show very large changes from the author ranked first to the one ranked second. Furthermore, these texts are attributed to their authors in all 80 of the tests based on more than the 50 most frequent words. Even Wharton and Chesnutt, in the middle of the graph, show a low Delta, a large change from rank 1 to rank 2, and consistent attribution over all 80 tests.

At the other end of the graph, the attribution of *The Titan* to Dreiser is quite weak: Delta is not very low, Delta-Z is low, both change very little from rank 1 to rank 2, and five different authors appear as likeliest in the 80 tests. Counterbalancing this somewhat, however, is the fact that Dreiser appears as the likeliest author in seventeen of the 24 tests involving the 600–800 most frequent words, and ranks no lower than

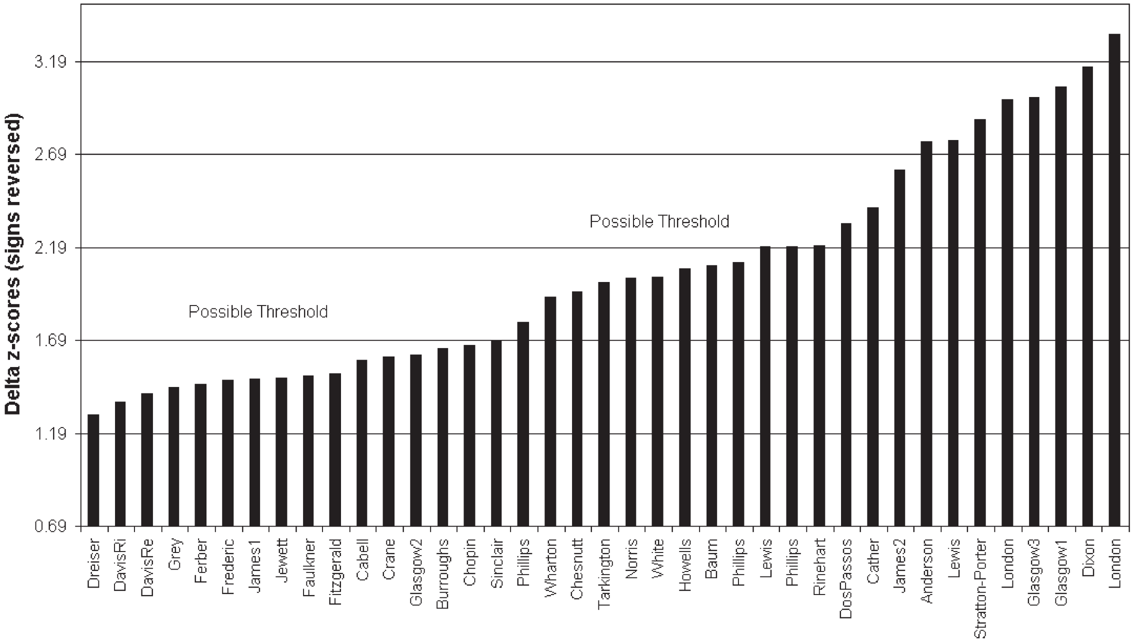

Fig. 6 Authorship simulation: members, errors, and others.

third (twice) in the remaining seven. Under these circumstances, one might be justified in suggesting Dreiser as a likely author and examining the matter further, in spite of the high Delta and low Delta-Z. Next to Dreiser in the graph, the texts by Rebecca Harding Davis and Richard Harding Davis are both incorrectly attributed to Wharton, but, had this been an actual attribution problem, Wharton could hardly have been taken seriously as the author of these novels. For *Frances Waldeaux,* Delta is high, Delta-Z is low, the change from rank 1 to rank 2 is very small, and seven different authors appear as likeliest in the 80 tests. For *Soldier of Fortune,* Delta is not as high, and the change from rank 1 to rank 2 for the Delta z- score is very large, but the change in Delta is less than 5%, and, again, seven different authors appear as likeliest. Furthermore, the likeliest authors change for both Davises even in tests involving the 600–800 most frequent words, where the results are most accurate for texts by members of the primary set.

For the two errors, James's *The Europeans*, and Glasgow's *Virginia*, with Delta-Z about the same as the Davises, matters are different: Delta is moderate, and the change from rank 1 to rank 2 for Delta-Z is substantial. Tarkington and Wharton each appear 12 times as the likeliest author for both of these novels in the 24 tests using the 600–800 most frequent words. The 12 Tarkington attributions come in tests with word lists that are not culled and the 12 Wharton attributions in tests with culled word lists. Given the greater overall accuracy of culled lists, it would have been reasonable to consider Wharton as the possible author of both of these texts, had it been a real authorship problem.

Table 1 Delta and Delta-Z for selected texts—analysis based on the 700 most frequent words, without personal pronouns and contractions, culled 70%

True Author / Novel	Likeliest Authors		Delta	% change Delta-Z	Diff. Attrib. in 80 tests
London	1	London	0.746	–3.335	1
White Fang	2	Twain	0.987	–0.777	
		% change	32%	–77%	
Dixon	1	Dixon	0.658	–3.159	1
The Leopard's Spots	2	Lewis	0.917	–0.994	
			39%	–69%	
Wharton	1	Wharton	0.614	–1.921	1
The House of Mirth	2	Howells	0.724	–1.186	
			18%	–38%	
Chesnutt	1	Chesnutt	0.660	–1.950	1
The Marrow of Tradition	2	Wharton	0.775	–1.149	
			17%	–41%	
Dreiser	1	Dreiser	0.820	–1.288	5
The Titan	2	Phillips	0.822	–1.272	
			0%	–1%	
DavisRe	1	Wharton	0.942	–1.405	7
Frances Waldeaux	2	Cather	0.949	–1.297	
			1%	–8%	
DavisRi	1	Wharton	0.830	–1.357	7
Soldiers of Fortune	2	Robins	0.871	–0.821	
			5%	–39%	
James	1	Wharton	0.777	–1.484	2
The Europeans	2	Tarkington	0.817	–1.182	
			5%	–20%	
Glasgow	1	Wharton	0.811	–1.610	6
Virginia	2	Tarkington	0.857	–1.139	
			6%	–29%	
Jewett	1	Jewett	0.924	–1.485	8
The Tory Lover	2	Wharton	0.936	–1.315	
			1%	–11%	
DosPassos	1	Glasgow	0.968	–2.319	1
Three Soldiers	2	Lewis	1.069	–1.434	
			10%	–38%	

Finally, consider the last two texts in Table 1. Jewett's Delta z-score is almost identical to James's, Delta is quite high, the change from rank 1 to rank 2 is small, and no fewer than eight different authors appear as likeliest in the 80 tests. Nevertheless, she is ranked as the likeliest author in 20 of the 24 tests with the 600–800 most frequent words. Finally, and most problematically, consider Dos Passos. The attribution of his text to Glasgow is very consistent, and is universal for the 80 analyses that include more than the 50 most frequent words. Delta is high, but Delta-Z is also very high, and both show substantial change from rank 1 to rank 2. Overall, the incorrect attribution of *Three Soldiers to Glasgow* seems stronger than the correct attribution of *The Tory Lover* to Jewett.

A full-fledged blind test on a similar group of authors and texts would be required to test the extent to which all of the factors examined here together lead to accurate attributions, but a preliminary test using a subset of the texts above produces cautionary results. To perform this test, 20 primary authors and texts were quickly selected, along with 20 texts by the same authors and 10 texts by authors not in the primary set. I wrote a simple program to rename the authors and texts of both sets and also the word frequency files for each of them. After putting the files away for a week (to reduce the effects of memory), I performed a series of forty Delta tests on them and then attempted to attribute the texts correctly based on Delta, Delta-Z, the change from rank 1 to rank 2, and the consistency of attribution. For this investigation, I tested the 100, 200, 300, 400, 500, 600, 700, and 800 most frequent words selected in five different ways:

1. The most frequent words
2. Personal pronouns removed
3. Culled at 70%
4. Culled at 70%; personal pronouns removed
5. Culled at 70%; contractions and personal pronouns removed

For 24 of the 30 test texts, all of the factors either suggest a firm attribution or indicate that the text is not by any of the authors of the primary set. Unfortunately, only 22 of these results prove correct, leaving two confident errors. Booth Tarkington's *Alice Adams* might be confidently attributed to an author other than those in the primary set because of a fairly high Delta, a low Delta-Z, very small changes from rank 1 to rank 2, and four different likeliest authors. (Tarkington appears as the likeliest author in 16 of the 30 tests, but Phillips is likeliest in 19.) As in the tests on 59 novels above, in this simulation, Dos Passos's *Three Soldiers* would seem safe to attribute to Glasgow because, in spite of a high Delta, Delta-Z is also high, there are large changes from rank 1 to rank 2, and the attribution to Glasgow is consistent across all 40 tests. Among the six less confident attributions, only one is correct. Although it would be unwise to generalize from this informal experiment, it does suggest that weak results should be treated with a great deal of skepticism, and even results that are very strong may occasionally be incorrect (reminding us that even statistically very unlikely events must sometimes happen). Furthermore, as Burrows suggests (2003, p. 24), false positive attributions seem more dangerous than false rejections, so that the case of Dos Passos is the most damaging one. Nevertheless, a blind procedure that correctly attributes 22 of 30 texts to their authors and produces only two clear errors should prove very useful as a first step in attribution in a

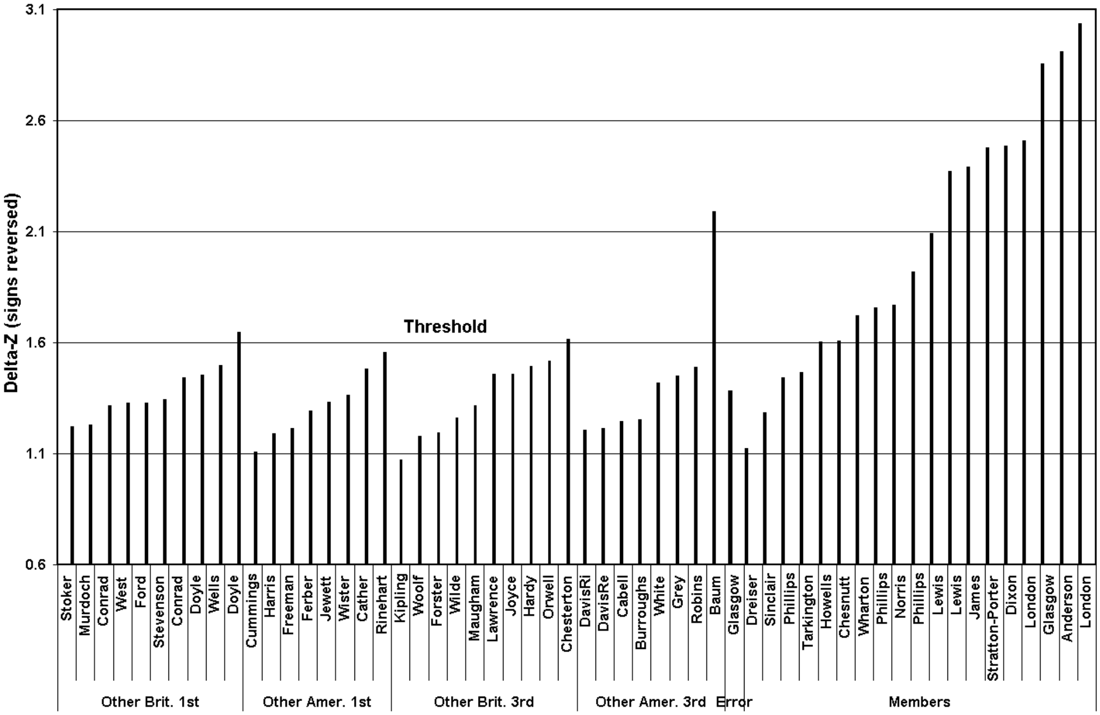

Fig. 7 Delta, nationality, and point of view: members, errors, and others of various kinds.

difficult open question of authorship. More precise and exhaustive methods should, in any case, uncover any spurious attributions.

6 Delta, Nationality, and Point of View

Now that the effectiveness of Delta for prose has been established, the effects of differences in the nationality of the authors and in the point of view of the narration can be examined. First, consider the results of tests on a larger set of British and American novelists. (Note that I am using the vexed term 'British' in an extremely loose and theoretically illegitimate way as shorthand for a clumsier but more accurate locution such as 'non-American novels in English'. For the present purposes, this seems unlikely to present serious difficulties.) These tests again involve sections of pure narration, with sections of 22 American third-person novels in the primary set.[10] The secondary set consists of 56 novels:

20 duplicate American third-person novels[11]
16 secondary American novels—8 first-person, 8 third-person[12]
20 secondary British novels—10 first-person, 10 third-person[13]

The results for the 700 most frequent words, culled at the 70% level, and with personal pronouns and contractions removed, are shown in Fig. 7. As can be seen, the only error among the twenty texts by authors in the primary set is again for Glasgow's *Virginia,* and all sets of secondary texts

[10] The texts are listed in the Appendix.
[11] The texts are listed in the Appendix.
[12] The texts are listed in the Appendix.
[13] The texts are listed in the Appendix.

seem to produce similar patterns. As in the analyses above, it is easier to attribute a duplicate text to the correct author than to eliminate the authors of the primary set as claimants. The results in Fig. 7 do not suggest that either nationality or point of view radically affects the results of Delta tests. If necessary, texts of different points of view or by authors of different nationalities can, with caution, be included in an authorship attribution problem. (The removal of personal pronouns from the analysis is especially important for any tests involving both first-person and third-person narration, for obvious reasons.)

A closer look at individual results suggests that nationality is less important than point of view: the difference between third-person and first-person secondary texts of the same nationality is generally greater than the difference between third-person American texts and third-person British texts or between first-person American and first-person British texts. This result may seem surprising given the well-known differences in British and American spelling, but a quick examination of the spelling of the texts shows that most of the different spellings are too infrequent to be significant, even when large numbers of frequent words are involved. For example, if the 800 most frequent words of the British and American novels are collected separately and compared, only a few seem to have radically different distributions: *toward* (American) versus *towards* (British), and *around* (American) versus *round* (British), are the only obvious discriminating pairs. In addition, *Negro, office, dollars, creek, ranch, judge,* and *store* are much more frequent in American texts, and *till, London,* and *Sir* are much more frequent in British texts. The well-known *-or* versus *-our* or *-er* versus *-re* spelling differences are too rare to have any effect: only *colour* appears among the 800 most frequent words in either the British or the American texts, ranking 775 in the British texts and 1106 in the American texts (*color* ranks 1021). None of the others appears frequently enough to be even potentially relevant. Removing the 14 distinctive words above has almost no effect on the accuracy of the analysis, so that they can be safely ignored. Although authorship questions involving authors of different nationalities might be expected to be rare, it is useful to know that Delta is a robust enough method to remain effective even in such difficult cases.[14]

7 Delta and a Large Group of Heterogeneous Texts

The accuracy of any authorship attribution technique, especially one based on a single measure, can be expected to decline as the number of texts increases. If the point of view of the texts and the nationalities of the authors also vary, we set the new method a very difficult task indeed. How well the method responds to such a task provides another measure of its robustness. Consider, then, a very large test, involving 102 texts by 59 authors. Here forty British and American first-person and third-person texts by 40 authors form the primary set:

[14] The difference in spelling between British and American novels has widened in the last century, so that this conclusion may not be valid for recent texts.

28 American novels—4 first-person, 24 third-person[15]
12 British novels—2 first-person, 10 third-person[16]

There are 62 texts in the secondary set:

30 American duplicates—5 first-person, 25 third-person[17]
13 British duplicates—4 first-person, 9 third-person[18]
10 American secondaries—all third-person[19]
9 British secondaries—4 first-person, 5 third-person[20]

In the two most accurate analyses, 34 of the 43 duplicates are correctly attributed. Some of the nine failures are, by now, predictable: neither James's *The Europeans* nor Glasgow's *Virginia* is properly attributed. Others, such as London's *The Sea Wolf,* Wells's *The Invisible Man,* and Kipling's *The Jungle Book,* have also been shown to be resistant to other authorship attribution tools (Hoover, 2001, 2002, 2003). Three others (as well as *The Sea Wolf)* are first-person narratives: Conrad's *Lord Jim,* Doyle's *The Valley of Fear*, and Stevenson's *Treasure Island.* Given the difficulty of this task, with its large number of texts that vary in both nationality and point of view, these results show that Delta is perhaps even more robust than might have been expected.

8 A Final Set of Tests

My final set of tests more closely resembles Burrows's tests, at least in scope, with a primary set of pure third-person authorial narration by 25 American authors[21] and a secondary set of 32 texts—16 by members[22] and 16 by others (nine American and seven British).[23] For this test, I have combined texts by a single author whenever there are more than two sections (as Burrows did with the poetry). This reduces the effects of variation among the author's texts. Delta tests on these texts produce what is by now a familiar pattern. When the 20–800 most frequent words are tested with pronouns removed, culled at 70% with pronouns removed, and culled at 70% with contractions and pronouns removed, there are six results in which 15 of the 16 texts by members of the primary set are attributed to the correct author. These are analyses using the 600, 700, and 800 most frequent words in the two culled sets just mentioned. In all six of these analyses, Delta fails for Upton Sinclair's *The Metropolis,* attributing it to Sinclair Lewis and ranking Upton Sinclair second. The pattern of Delta-Z for members, error, and others, shown in Fig. 8, is also familiar, with three texts by others rising above the threshold and three texts by members (in addition to the error) falling below it. What makes this final graph particularly interesting, however, is the very strong results for Glasgow's *The Battle Ground*, James's *The Wings of the Dove,* and Phillips's *The Fashionable Adventures of Joshua Craig*, in spite of the fact that these authors' problematic novels, *Virginia, The Europeans,* and *Susan Lenox* are included in the combined texts for these authors in the primary set.

[15] The texts are listed in the Appendix.
[16] The texts are listed in the Appendix.
[17] The texts are listed in the Appendix.
[18] The texts are listed in the Appendix.
[19] The texts are listed in the Appendix.
[20] The texts are listed in the Appendix.
[21] The texts are listed in the Appendix.
[22] The texts are listed in the Appendix.
[23] The texts are listed in the Appendix.

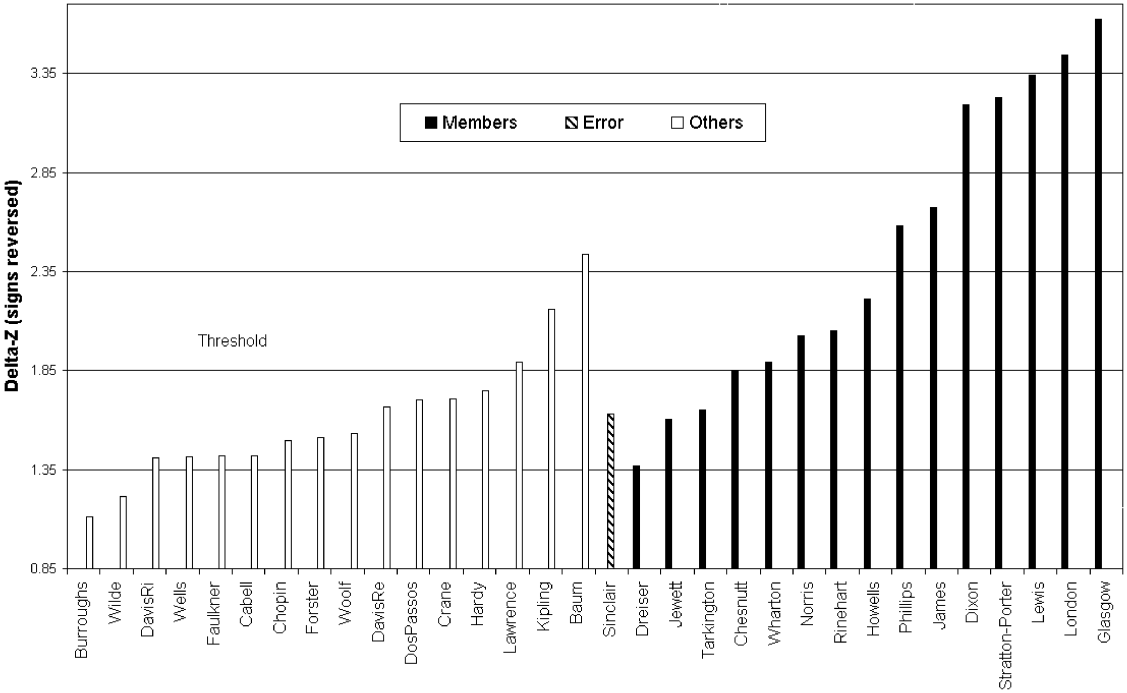

Fig. 8 Members, errors, and others for 57 novels.

9 Conclusion

The results above show that, as Burrows found with Restoration poetry, Delta does a very good job of attributing secondary prose texts to the correct primary authors and rejecting the primary authors as claimants for secondary texts by other authors. They also confirm Burrows's conclusion that the accuracy of the tests increases as the number of frequent words increases from 40 to 150 and extend this conclusion by showing that the results continue to increase in accuracy when much larger numbers of frequent words are analyzed. Although the traditional view has been that only the most frequent words, typically function words, are likely to be beyond the author's control and are therefore suitable for authorship attribution, the analyses above show that Delta's accuracy continues to increase, often substantially, as the word frequency lists increase, at least up to the 600 or 700 most frequent words, at which point almost all the words are content words. This surprising result needs further investigation.

Testing large numbers of differently selected word frequency lists shows that removing personal pronouns and words for which a single text supplies more than 70% of the occurrences greatly increases the accuracy of the analyses, and that removing contractions sometimes increases, but more often decreases, their accuracy. The results above for prose are generally less accurate than those Burrows achieved for poetry, but they are still quite robust, suggesting that Delta will be very useful in open attribution problems on prose as well as poetry.

The results above confirm Burrows's observation that a small Delta, a large Delta-Z, and consistency of attribution across various numbers of frequent words are indicators of reliable attribution. They also suggest that large changes in Delta and Delta-Z from the likeliest to the second likeliest author are characteristic of reliability. Even when all of these indicators agree, authorship simulations show that false attributions remain a possibility, but Delta is robust enough to be tested further on texts of differing points of view and different nationalities, where it continues to perform well. Testing Delta on a very large and diverse set of texts shows that it does a surprisingly good job of attributing texts to their correct authors and in rejecting authors in the primary set as the authors of texts by other authors. Finally, the results above suggest that creating samples that combine several texts for each of the authors in the primary set of texts, as Burrows did, helps to improve accuracy by limiting the effect of stylistic variation within an author's works. Fortunately, this technique will generally also be possible in real authorship attribution problems.

Further research may discover additional ways of limiting the likelihood of attribution errors. Even without such improvements, however, Delta should be very effective in reducing the number of potential claimants and suggesting likely authors in some of the most difficult problems of authorship attribution, allowing the analyst to focus more exact and powerful, but also more time-consuming, methods on a much smaller group of authors. Even if no purely statistical technique ever proves sufficient for correct attribution of all texts and all authors, Burrows's Delta is a valuable addition to the authorship attribution toolbox.

References

Burrows, J. F. (2001). *Questions of Authorship: Attribution and Beyond. Association for Computers and the Humanities and Association for Literary and Linguistic Computing, Joint International Conference*, New York, 14 June 2001.

Burrows, J. F. (2002a). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**: 267–87.

Burrows, J. F. (2002b). The Englishing of Juvenal: computational stylistics and translated texts. *Style*, **36**: 677–99.

Burrows, J. F. (2003). Questions of authorship: attribution and beyond. *Computers and the Humanities,* **37**: 5–32.

Hoover, D. L. (2001). Statistical stylistics and authorship attribution: an empirical investigation. *Literary and Linguistic Computing*, **16**: 421–44.

Hoover, D. L. (2002). Frequent word sequences and statistical stylistics. *Literary and Linguistic Computing,* **17**: 157–80.

Hoover, D. L. (2003). Frequent collocations and authorial style. *Literary and Linguistic Computing,* **18**: 261–86.

Richards, M. K. (1971). *Ellen Glasgow's Development as a Novelist.* The Hague: Mouton.

Appendix

2 The texts in the primary set are as follows: *Winesburg, Ohio,* Anderson (1919) 38,051 words; *The Professor's House,* Cather (1925) 21,871; *The House Behind the Cedars,* Chesnutt (1900) 27,049; *The Clansman,* Dixon (1905) 26,543; *Sister Carrie,* Dreiser (1900) 36,269; *The Battle-Ground,* Glasgow (1902) 27,359; *A Hazard of New Fortunes,* Howells (1890) 27,187; *The Ambassadors,* James (1909) 32,664; *The Country Doctor*, Jewett (1884) 25,551; *Main Street,* Lewis (1920) 31,006; *The Call of the Wild,* London (1903) 27,602; *McTeague,* Norris (1899) 30,794; *The Conflict,* Phillips (1911) 29,612; *The Breaking Point,* Rinehart (1922) 22,558; *The Mills of the Gods*, Robins (1920) 10,562; *The Jungle,* Sinclair (1906) 28,656; *Freckles*, Stratton-Porter (1904) 24,536; *Alice Adams,* Tarkington (1921) 20,763; *The Tragedy of Pudd'nhead Wilson,* Twain (1894) 23,489; *The Age of Innocence,* Wharton (1920) 26,967.

5 The 25 texts by authors in the primary set are *Marching Men,* Anderson (1917) 28,756 words; *Song of the Lark,* Cather (1915) 29,394; *The Marrow of Tradition*, Chesnutt (1901) 33,447; *The Leopard's Spots,* Dixon (1902) 20,972; *The Titan,* Dreiser (1914) 27,082; *The Deliverance,* Glasgow (1904) 24,014; *Virginia,* Glasgow (1913) 23,763; *Voice of the People*, Glasgow (1900) 33,236; *The Kentons,* Howells (1890) 34,282; *The Europeans,* James (1878) 24,511; *The Wings of the Dove,* James (1909) 34,526; *The Tory Lover,* Jewett (1901) 22,337; *Babbitt,* Lewis (1922) 26,033; *Our Mr. Wrenn,* Lewis (1914) 19,236; *Burning Daylight,* London (1910) 28,298; *White Fang,* London (1906) 25,877; *Octopus,* Norris (1901) 32,314; *The Fashionable Adventures of Joshua Craig*, Phillips (1909) 21,954; *The Grain of Dust*, Phillips (1911) 23,492; *Susan Lenox,* Phillips (1917) 27,739; *Dangerous Days,* Rinehart (1919) 21,991; *The Metropolis,* Sinclair (1908) 27,514; *A Girl of the Limberlost*, Stratton-Porter (1909) 25,027; *The Conquest of Canaan*, Tarkington (1905) 14,821; *The House of Mirth,* Wharton (1905) 36,826. The fourteen texts by authors not in the main set are *The Wonderful Wizard of Oz,* Baum (1900) 17,425; *Tarzan of the Apes*, Burroughs (1914) 38,849; *Jurgen: a Comedy of Justice*, Cabell (1919) 22,077; *The Awakening,* Chopin (1899) 21,429; *The Red Badge of Courage*, Crane (1895) 33,898; *Frances Waldeaux*, Davis (1897) 10,311; *Soldiers of Fortune*, Davis (1897) 34,834; *Three Soldiers*, Dos Passos (1921) 27,532; *Light in August*, Faulkner (1932) 32,639; *Emma McChesney & Co.*, Ferber (1915) 16,969; *This Side of Paradise*, Fitzgerald (1920) 17,524; *The Damnation of Theron Ware*, Frederic (1896) 35,562; *The Man of the Forest*, Grey (1919) 29,319; *The Silent Places*, White (1904) 21,509.

10 *Winesburg, Ohio,* Anderson (1919) 38,051 words; *The House Behind the Cedars,* Chesnutt (1900) 27,049; *The Awakening,* Chopin (1899) 21,429; *The Red Badge of Courage*, Crane (1895) 33,898; *The Leopard's Spots,* Dixon (1902) 20,972; *Three Soldiers*, Dos Passos (1921) 27,532; *Sister Carrie,* Dreiser (1900) 36,269; *Light in August*, Faulkner (1932) 32,639; *This Side of Paradise*, Fitzgerald (1920) 17,524; *The Damnation of Theron Ware*, Frederic (1896) 35,562; *The Battle-Ground,* Glasgow (1902) 27,359; *The Kentons,* Howells (1890) 34,282; *The Ambassadors,* James (1909) 32,664; *Main Street,* Lewis (1920) 31,006; *The Call of the Wild,* London (1903) 27,602; *McTeague,* Norris (1899) 30,794; *The Conflict,* Phillips (1911) 29,612; *The Jungle,* Sinclair (1906) 28,656; *Freckles*, Stratton-Porter (1904) 24,536; *Alice Adams,* Tarkington (1921) 20,763; *The Tragedy of Pudd'nhead Wilson,* Twain (1894) 23,489; *The Age of Innocence,* Wharton (1920) 26,967.

11 *Marching Men,* Anderson (1917) 28,756 words; *The Marrow of Tradition*, Chesnutt (1901) 33,447; *The Clansman,* Dixon (1905) 26,543; *The Titan,*

Dreiser (1914) 27,082; *The Deliverance,* Glasgow (1904) 24,014; *Virginia,* Glasgow (1913) 23,763; *A Hazard of New Fortunes,* Howells (1890) 27,187; *The Wings of the Dove,* James (1909) 34,526; *Babbitt,* Lewis (1922) 26,033; *Our Mr. Wrenn,* Lewis (1914) 19,236; *Burning Daylight,* London (1910) 28,298; *White Fang,* London (1906) 25,877; *Octopus,* Norris (1901) 32,314; *The Fashionable Adventures of Joshua Craig*, Phillips (1909) 21,954; *The Grain of Dust*, Phillips (1911) 23,492; *Susan Lenox,* Phillips (1917) 27,739; *The Metropolis,* Sinclair (1908) 27,514; *A Girl of the Limberlost*, Stratton-Porter (1909) 25,027; *The Conquest of Canaan*, Tarkington (1905) 14,821; *The House of Mirth,* Wharton (1905) 36,826.

12 First-person: *My Antonia,* Cather (1918) 23,841 words; *The Enormous Room*, Cummings (1922) 28,184; *Dawn O'Hara, the Girl who Laughed*, Ferber (1911) 24,750; *The Heart's Highway*, Freeman (1900) 25,780; *The Bomb*, Harris (1908) 22,753; *The Country of the Pointed Firs*, Jewett (1896) 18,935; *The Case of Jennie Brice*, Rinehart (1913) 18,034; *The Virginian*, Wister (1902) 22,470. Third-person: *The Wonderful Wizard of Oz,* Baum (1900) 17,425; *Tarzan of the Apes*, Burroughs (1922) 38,849; *Jurgen: a Comedy of Justice*, Cabell (1922) 22,077; *Frances Waldeaux*, Davis (1897) 10,311; *Soldiers of Fortune*, Davis (1905) 34,834; *The Man of the Forest*, Grey (1920) 29,319; *The Mills of the Gods*, Robins (1920) 10,562; *The Silent Places*, White (1904) 21,509.

13 First-person: *The Nigger of the Narcissus*, Conrad (1897) 24,094 words; *The Hound of the Baskervilles*, Doyle (1901–02) 12,073; *The Valley of Fear,* Doyle (1914–15) 16,775; *The Good Soldier*, Ford (1915) 34,443; *Under the Net*, Murdoch (1954) 12,746; *Treasure Island*, Stevenson (1883) 11,055; *Dracula*, Stoker (1897) 17,242; *The War of the Worlds*, Wells (1898) 26,621; *The Return of The Soldier*, West (1918) 15,706. Third-person: *The Man Who Was Thursday*, Chesterton (1908) 24,612; *Howards End*, Forster (1910) 18,744; *Jude the Obscure*, Hardy (1896) 20,616; *A Portrait of the Artist as a Young Man*, Joyce (1916) 13,122; *The Light That Failed*, Kipling (1890) 19,428; *Sons and Lovers*, Lawrence (1913) 17,404; *Of Human Bondage*, Maugham (1915) 22,088; *Nineteen Eighty-Four*, Orwell (1949) 25,638; *The Picture of Dorian Gray*, Wilde (1891) 17,724; *To the Lighthouse*, Woolf (1927) 15,054.

15 American: *Winesburg, Ohio,* Anderson (1919) 38,051 words; *The Professor's House,* Cather (1925) 21,871; *The House Behind the Cedars,* Chesnutt (1900) 27,049; *The Red Badge of Courage*, Crane (1895) 33,898; *The Enormous Room*, Cummings (1922) 28,184; *Soldiers of Fortune*, Davis (1905) 34,834; *The Leopard's Spots*, Dixon (1902) 20,972; *Sister Carrie,* Dreiser (1900) 36,269; *Light in August*, Faulkner (1932) 32,639; *Emma McChesney & Co.*, Ferber (1915) 16,969; *The Damnation of Theron Ware*, Frederic (1896) 35,562; *The Heart's Highway*, Freeman (1900) 25,780; *The Battle-Ground,* Glasgow (1902) 27,359; *The Bomb*, Harris (1908) 22,753; *The Kentons,* Howells (1890) 34,282; *The Ambassadors,* James (1909) 32,664; *The Country of the Pointed Firs*, Jewett (1896) 18,935; *Main Street,* Lewis (1920) 31,006; *The Call of the Wild,* London (1903) 27,602; *McTeague,* Norris (1899) 30,794; *The Conflict,* Phillips (1911) 29,612; *The Breaking Point,* Rinehart (1922) 22,558; *The Mills of the Gods*, Robins (1920) 10,562; *The Jungle,* Sinclair (1906) 28,656; *Freckles*, Stratton-Porter (1904) 24,536; *Alice Adams,* Tarkington (1921) 20,763; *The Age of Innocence,* Wharton (1920) 26,967; *The Virginian*, Wister (1902) 22,470.

16 *The Nigger of the Narcissus,* Conrad (1897) 24,094 words; *The Hound of the Baskervilles*, Doyle (1901–02) 12,073; *Howards End*, Forster (1910) 18,744; *Jude the Obscure*, Hardy (1896) 20,616; *Brave New World*, Huxley (1932) 9,951; *A Portrait of the Artist as a Young Man*, Joyce (1916) 13,122; *The Light*

That Failed, Kipling (1890) 19,428; *Lady Chatterley's Lover*, Lawrence (1928) 23,931; *Nineteen Eighty-Four*, Orwell (1949) 25,638; *The Black Arrow*, Stevenson (1888) 22,499; *When the Sleeper Wakes*, Wells (1899) 25,065; *The Voyage Out,* Woolf (1915) 26,386.

17 *Marching Men,* Anderson (1917) 28,756 words; *My Antonia,* Cather (1918) 23,841; *Song of the Lark,* Cather (1915) 29,394; *The Marrow of Tradition*, Chesnutt (1901) 33,447; *The Clansman,* Dixon (1905) 26,543; *The Titan,* Dreiser (1914) 27,082; *Dawn O'Hara, the Girl Who Laughed*, Ferber (1911) 24,750; *The Deliverance,* Glasgow (1904) 24,014; *Virginia,* Glasgow (1913) 23,763; *Voice of the People*, Glasgow (1900) 33,236; *A Hazard of New Fortunes,* Howells (1890) 27,187; *The Europeans,* James (1878) 24,511; *The Wings of the Dove,* James (1909) 34,526; *The Country of the Pointed Firs*, Jewett (1896) 18,935; *The Tory Lover,* Jewett (1901) 22,337; *Babbitt,* Lewis (1922) 26,033; *Our Mr. Wrenn,* Lewis (1914) 19,236; *Burning Daylight,* London (1910) 28,298; *The Sea Wolf,* London (1904) 22,016; *White Fang,* London (1906) 25,877; *Octopus,* Norris (1901) 32,314; *The Fashionable Adventures of Joshua Craig*, Phillips (1909) 21,954; *The Grain of Dust*, Phillips (1911) 23,492; *Susan Lenox,* Phillips (1917) 27,739; *The Case of Jennie Brice*, Rinehart (1913) 18,034; *Dangerous Days,* Rinehart (1919) 21,991; *The Metropolis,* Sinclair (1908) 27,514; *A Girl of the Limberlost*, Stratton-Porter (1909) 25,027; *The Conquest of Canaan*, Tarkington (1905) 14,821; *The House of Mirth,* Wharton (1905) 36,826.

18 *Lord Jim,* Conrad (1900) 28,171 words; *The Valley of Fear*, Doyle (1914–15) 16,775; *Room with a View*, Forster (1908) 24,357; *The Mayor of Casterbridge*, Hardy (1922) 17,064; *Tess of the d'Urbervilles,* Hardy (1891) 15,175; *The Jungle Book,* Kipling (1893) 10,002; *Kim*, Kipling (1901) 16,407; *Stalky & Co,* Kipling *(*1899) 14,398; *Sons and Lovers*, Lawrence (1913) 17,404; *Treasure Island,* Stevenson (1883) 11,055; *The Invisible Man,* Wells (1897) 14,072; *The War of the Worlds*, Wells (1898) 26,621; *To the Lighthouse*, Woolf (1927) 15,054.

19 *The Wonderful Wizard of Oz,* Baum (1900) 17,425 words; *Tarzan of the Apes*, Burroughs (1914) 38,849; *Jurgen: a Comedy of Justice*, Cabell (1919) 22,077; *The Awakening,* Chopin (1899) 21,429; *Frances Waldeaux*, Davis (1897) 10,311; *Three Soldiers*, Dos Passos (1921) 27,532; *This Side of Paradise*, Fitzgerald (1920) 17,524; *The Man of the Forest*, Grey (1919) 29,319; *The Tragedy of Pudd'nhead Wilson,* Twain (1894) 23,489; *The Silent Places*, White (1904) 21,509.

20 [*The Man Who Was Thursday*, Chesterton (1908) 24,612 words; *The Good Soldier*, Ford (1915) 34,443; *The Inheritors,* Golding (1954) 17,457; *A Single Man,* Isherwood (1964) 11,282; *Of Human Bondage*, Maugham (1915) 22,088; *Under the Net*, Murdoch (1954) 12,746; *Dracula*, Stoker (1897) 17,242; *The Return of the Soldier*, West (1918) 15,706; *The Picture of Dorian Gray*, Wilde (1891) 17,724.

21 *Marching Men + Winesburg, Ohio,* Anderson (1917, 1919) 66,807 words; *The Professor's House + Song of the Lark,* Cather (1925, 1915) 51,265; *The House Behind the Cedars,* Chesnutt (1900) 27,049; *The Clansman,* Dixon (1905) 26,543; *Sister Carrie,* Dreiser (1900) 36,269; *Emma McChesney & Co.*, Ferber (1915) 16,969; *This Side of Paradise*, Fitzgerald (1920) 17,524; *The Damnation of Theron Ware*, Frederic (1896) 35,562; *The Deliverance + Virginia + Voice of the People,* Glasgow (1913, 1904, 1900) 81,013; *The Man of the Forest*, Grey (1919) 29,319; *A Hazard of New Fortunes,* Howells (1890) 27,187; *The Ambassadors + The Europeans*, James (1909, 1878) 57,175; *The Country*

Doctor, Jewett (1884) 25,551; *Babbitt + Our Mr. Wrenn*, Lewis (1922, 1914) 45,269; *Burning Daylight + White Fang*, London (1910, 1906) 54,175; *McTeague,* Norris (1899) 30,794; *The Conflict + The Grain of Dust + Susan Lenox*, Phillips (1911, 1911, 1917) 80,843; *The Breaking Point,* Rinehart (1922) 22,558; *The Mills of the Gods*, Robins (1920) 10,562; *The Jungle,* Sinclair (1906) 28,656; *Freckles*, Stratton-Porter (1904) 24,536; *Alice Adams,* Tarkington (1921) 20,763; *The Tragedy of Pudd'nhead Wilson,* Twain (1894) 23,489; *The Age of Innocence,* Wharton (1920) 26,967; *The Silent Places*, White (1904) 21,509.

22 *The Marrow of Tradition*, Chesnutt (1901) 33,447 words; *The Leopard's Spots,* Dixon (1902) 20,972; *The Titan,* Dreiser (1914) 27,082; *The Battle-Ground,* Glasgow (1902) 27,359; *The Kentons,* Howells (1890) 34,28; *The Wings of the Dove,* James (1909) 34,526; *The Tory Lover,* Jewett (1901) 22,337; *Main Street,* Lewis (1920) 31,006; *The Call of the Wild,* London (1903) 27,602; *Octopus,* Norris (1901) 32,314; *The Fashionable Adventures of Joshua Craig*, Phillips (1909) 21,954; *Dangerous Days,* Rinehart (1919) 21,991; *The Metropolis* (1908) 27,514; *A Girl of the Limberlost*, Stratton-Porter (1909) 25,027; *The Conquest of Canaan*, Tarkington (1905) 14,821; *The House of Mirth,* Wharton (1905) 36,826.

23 *The Wonderful Wizard of Oz,* Baum (1900) 17,425 words; *Tarzan of the Apes*, Burroughs (1914) 38,849; *Jurgen: a Comedy of Justice*, Cabell (1919) 22,077; *The Awakening,* Chopin (1899) 21,429; *The Red Badge of Courage*, Crane (1895) 33,898; *Frances Waldeaux*, Davis (1897) 10,311; *Soldiers of Fortune*, Davis (1905) 34,834; *Three Soldiers*, Dos Passos (1921) 27,532; *Light in August*, Faulkner (1932) 32,639; *Howards End + Room with a View*, Forster (1910, 1908) 43,101; *Jude the Obscure + The Mayor of Casterbridge + Tess of the d'Urbervilles*, Hardy (1896, 1922, 1891) 52,855; *The Jungle Book +Kim + The Light That Failed + Stalky & Co*, Kipling (1893, 1901, 1899, 1890) 60,235; *Lady Chatterley's Lover + Sons and Lovers*, Lawrence (1928, 1913) 41,335; *The Invisible Man + When the Sleeper Wakes*, Wells (1897, 1899) 39,137; *The Picture of Dorian Gray,* Wilde (1891) 17,724; *To the Lighthouse + The Voyage Out*, Woolf (1927, 1915) 41,440.