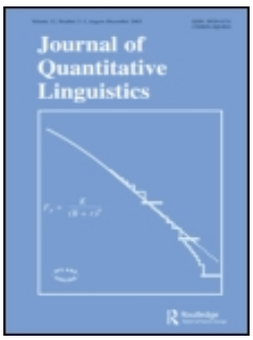

Routledge
Taylor & Francis Group

Journal of Quantitative Linguistics

ISSN: 0929-6174 (Print) 1744-5035 (Online) Journal homepage: https://www.tandfonline.com/loi/njql20

Improving Authorship Attribution: Optimizing Burrows' Delta Method*

Peter W. H. Smith & W. Aldridge

To cite this article: Peter W. H. Smith & W. Aldridge (2011) Improving Authorship Attribution: Optimizing Burrows' Delta Method*, Journal of Quantitative Linguistics, 18:1, 63-88, DOI: 10.1080/09296174.2011.533591

To link to this article: https://doi.org/10.1080/09296174.2011.533591

Journal of Quantitative Linguistics
2011, Volume 18, Number 1, pp. 63–88
DOI: 10.1080/09296174.2011.533591

Routledge
Taylor & Francis Group

Improving Authorship Attribution: Optimizing Burrows' Delta Method*

Peter W. H. Smith[1] and W. Aldridge[2]
[1]City University London; [2]E-Dialog, London

ABSTRACT

Burrows' Delta Method (Burrows, 2002) is a leading method of authorship attribution. It can be used to shortlist potential authors from a list or to even identify potential authors. The technique has been extended by Hoover (2004a, 2006). In this investigation, we look at the choice of words for the word vector used, the size of the word vector, the similarity measure and the impact of corpus choice on the accuracy of text classification. Our results show a word frequency vector of between 200 and 300 words give the most accurate results (Aldridge, 2007). We also demonstrate a dramatic improvement in accuracy by adapting Burrows' Delta to the cosine similarity measure. Additionally, our results indicate areas where the word vector can be optimized still further for more accurate results.

INTRODUCTION

Burrows (2002) identified that techniques for authorship attribution were suitable only within a ''closed game'', i.e. the author under suspicion was selected from a small list (sometimes, only one or two suspect authors). Burrows identified the need to provide a means by which an unattributed text may be compared with a large set of authors to establish a potential author, or at least a shortlist of authors. He proposed his Delta technique for this purpose. This raises the issue in more general terms of multiple versus two-way classification.

*Address correspondence to: Peter W.H. Smith, Department of Computing, City University, Northampton Square, London, EC1V 0HB, England.
Tel: +44(0)20-70408437. E-mail: peters@soi.city.ac.uk

The Delta technique works by computing the sum of a set of *z*-scores for an unknown text against a large text corpus of a contemporary period. The *z*-scores are then summed to compute the delta value. Delta values are also computed for a series of texts by known authors. The author whose text has the lowest delta value is then attributed as the author of the unknown work. A fully worked example is given in Burrows (2002). It is a powerful technique because it is simple and may be applied flexibly. It expresses the difference between two texts as a single measurement; the larger it is, the more dissimilar the comparative texts are to each other. The method uses most frequent word profiles. Burrows demonstrates that the method can be effective in producing candidate shortlists for texts as little as 100 words long, but for authorship attribution, texts of at least 1500 words are required. A fully worked example is given in Figure 1. As the number of words used in the word frequency vector is increased, the method shows a slowly increasing accuracy (Figure 2).

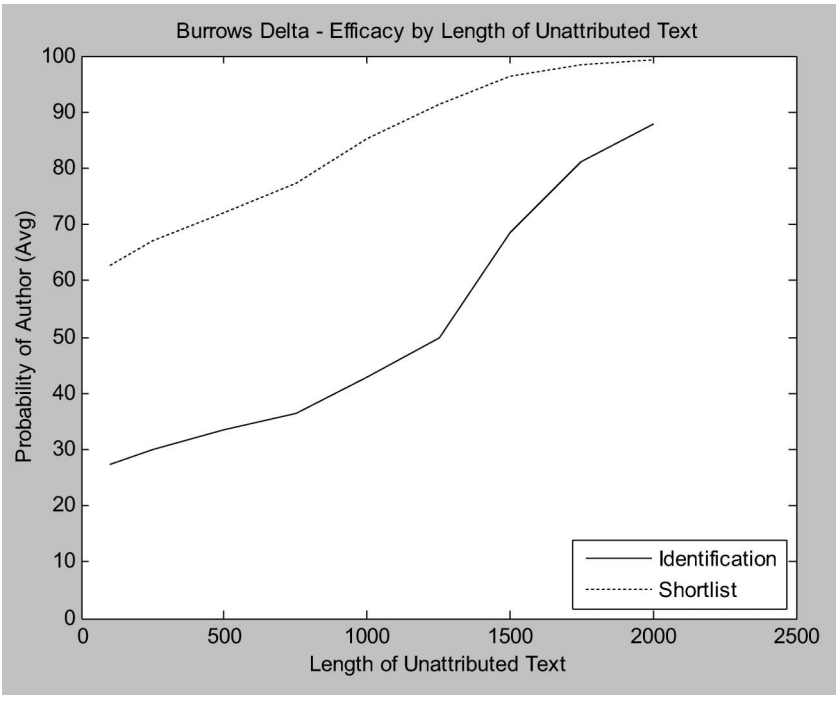

Fig. 1. Efficacy by length of text.

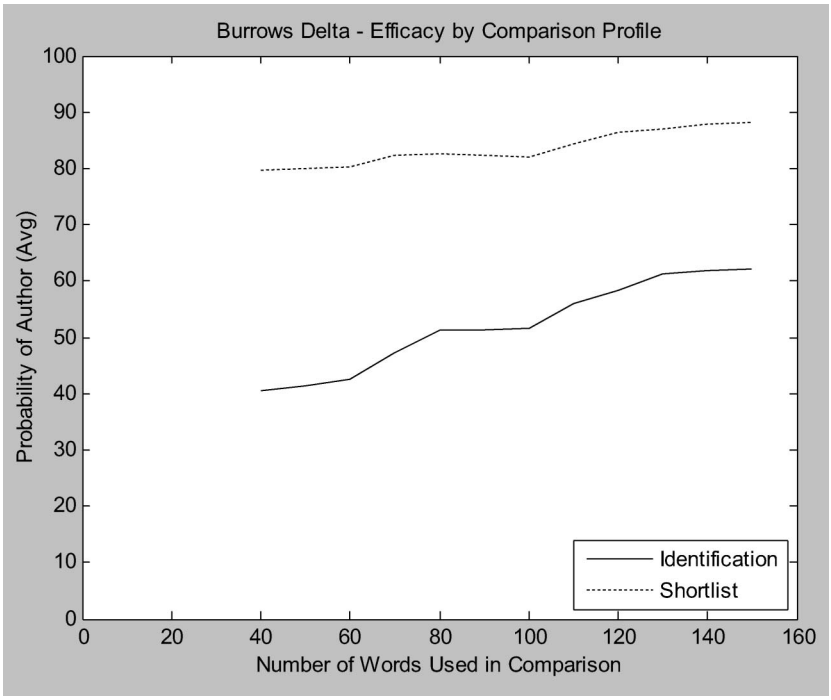

Fig. 2. Efficacy by word vector size.

Hoover (2004a) demonstrated that Burrows' Delta Measure was almost as effective when applied to prose, testing it on a set of American novels. However, in his experiments, unlike Burrows, Hoover made no distinctions between parts of speech. Both Hoover and Burrows reported improved performance when the word frequency vector is increased (Figure 3). The Hoover experiment worked significantly better on author identification and Burrows performs better on author short-listing. Hoover then identified that improvements continue above the 150 word vector size where Burrows' experiment stopped (Figure 4). Figure 4 should be considered in three sections: firstly up to 200 words which show a steady improvement in performance, with the exception of a possible anomaly at around 50 words. Secondly, between 200 and 500 words there is no improvement in author short-listing but the ability to identify the correct author slowly improves. Above 500 words, there is a levelling off of both short-listing and authorship attribution. It is probably worth noting that the

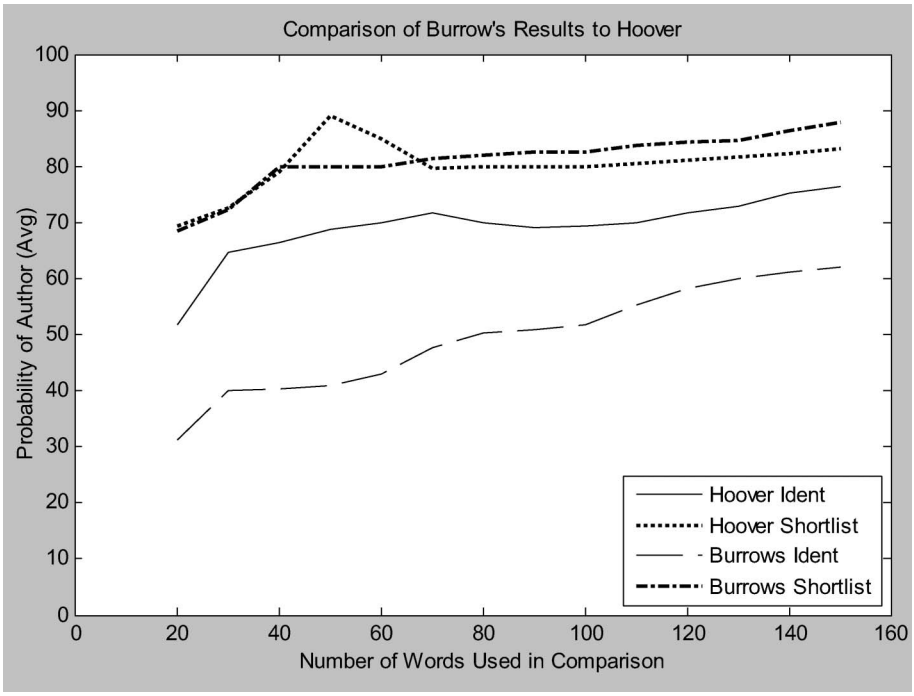

Fig. 3. A comparison of Burrows' and Hoover's results.

use of such a large word vector greatly increases the chances of comparisons based on hapaxes within the comparison vector.

Hoover (2004a) also identified some improvement by excluding personal pronouns from the word vector and discovered that excluding contracted words (e.g. couldn't, ain't, etc.) made the performance worse. He additionally observed that accuracy was very slightly improved by including only words with a positive delta score, but was significantly worse if only negative *z*-scores were chosen. Hoover experimented further by adding weightings and being selective with *z*-scores, which he called Delta Primes, although these were all Euclidean distance measures and had the effect of discriminating in favour of positive *z*-scores. Hoover (2006) extended the word vector still further to cover the entire set of words available in any test document. As lower frequency words were primarily content words and as these also had mostly positive *z*-scores, he concluded that content words must be responsible for the improvements in the accuracy of the Delta method. The further claim that content

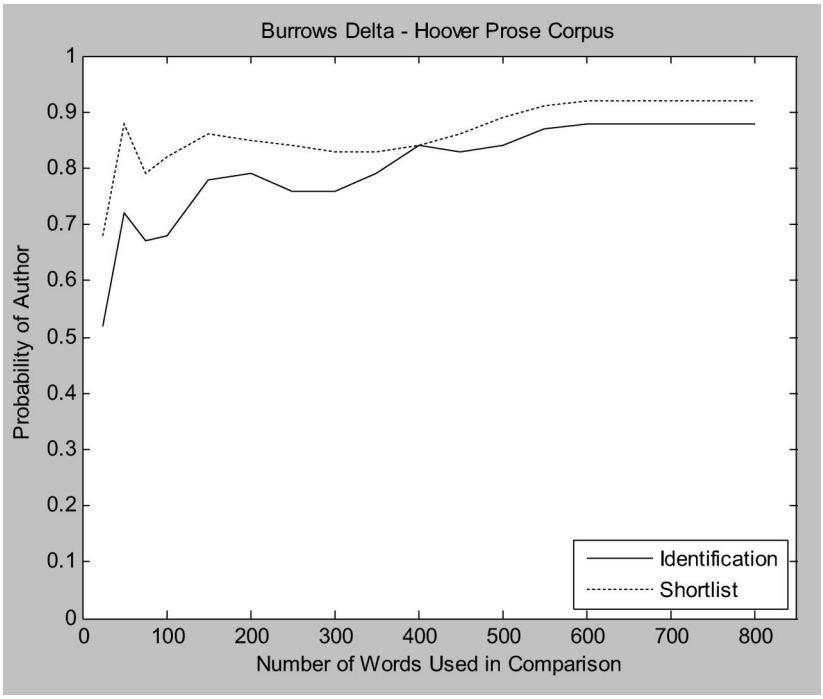

Fig. 4. Use of larger word vectors.

words were more important for authorship attribution was largely unsubstantiated.

There are also other concerns about the Delta Method that need to be addressed. Firstly, it uses a Euclidean distance measure, which is known to become inaccurate as dimensionality increases (Frigui & Nasraoui, 2003). Secondly, it makes the assumption that words conform to a normal distribution and there have to be significant concerns about this for lower frequency words. Mosteller and Wallace (1964), for example used a Poisson distribution. In Parbat (2008) a gamma distribution was used in place of the assumption of normality for low frequency words and the results were promising. The use of a gamma/Poisson model seems worth investigating further. Also, there is no totally objective way of determining whether a disputed text is by one of the authors included in the test. The author with the lowest score becomes the attributed author, but how do we determine when there is a completely negative result? The word choice for the Delta method is based purely on

frequency of occurrence in the corpus and there is no objective way of determining how many words should be chosen. Particularly because of the curse of dimensionality, we should not assume that simply adding more words to the word frequency vector will result in better classification results. Although the Delta method demonstrates a method of multi-way classification, there have to be some doubts as to whether multi-way classification is more accurate than a series of two-way classification tests. This also raises the issue of non-classification, i.e. producing results that indicate that none of the chosen authors are the author of the disputed text.

METHOD

Experiments were undertaken on data prepared from a corpus of 17th-century poetry by 25 authors from the English restoration period (Edmond Waller, 1606–1687 to William Congreve, 1670–172), following as closely as possible to Burrows' original corpus and methodology. A selection of the corpus was reserved for testing. The training set was also subdivided into 10 sets of equal sizes and each experiment was repeated 10 times, each time leaving out one of the subsets. Each test measured accuracy of classification based on identifying the author correctly and of producing a shortlist of five that contain the correct author. We then applied a *t*-test to test for significance of the results.

Two corpora of poetry were compiled – the Stuart Corpus comprising of 581,768 words from 1068 poems, the work of 24 English poets. This corpus was extended to comprise a total of 195 poems in total, 954,122 words. The corpus was checked for accuracy and provenance. The poems were taken from Oldpoetry (www.oldpoetry.com) and were checked against texts at Project Gutenberg (see www.gutenberg.org) and *New Oxford Book of English Verse* (Gardner, 1972).

The principal areas for investigation in this study are:

- An evaluation of word frequency vector size. Experiments use words vectors of 50, 100, 150, 200, 300, 400, 500, 600, 700 and 800 of the most frequently occurring words in the corpus.
- An evaluation of text length and the accuracy of Burrows' Delta. Experiments measure the performance on poems of up to 50 words,

50–100 words, 100–500 words, 500–1000 words, 1000–2000 words and >2000 words.

- An evaluation of the composition of the word vector using the following word frequency vectors:

(1) The top N most frequently occurring words in the corpus.
(2) The top N + offset most frequently occurring words, where offset is a variable parameter.
(3) The top N odd numbered words by frequency.
(4) The most frequent words in the corpus, placing function words above content words.
(5) The top N most frequent content words.
(6) A random selection from the top 1600 words.
(7) Words ordered on word length, taking smallest words first.
(8) The N least frequently occurring words in the corpus.

- An evaluation of the distance measure, comparing a Euclidean distance measure with an angular similarity measure.
- An evaluation of the importance of compiling the most appropriate corpus to be used for comparison.

RESULTS

Direct Comparison with Burrows' Original Results

The first part of the experiment involved carrying out a direct comparison with Burrows' original results. We kept our corpus as close as possible to the one used by Burrows. Tests were carried out evaluating the Delta method against poem length; Figures 5 and 6 show broad agreement with Burrows' original results. This provides the foundation and justification for the results that follow.

The Impact of Text Length

We tested the impact of text length on the accuracy of Burrows' Delta. Each poem was tested against slightly different text corpuses as described in Section 1. The results are shown as box-plots in Figures 7 and 8.

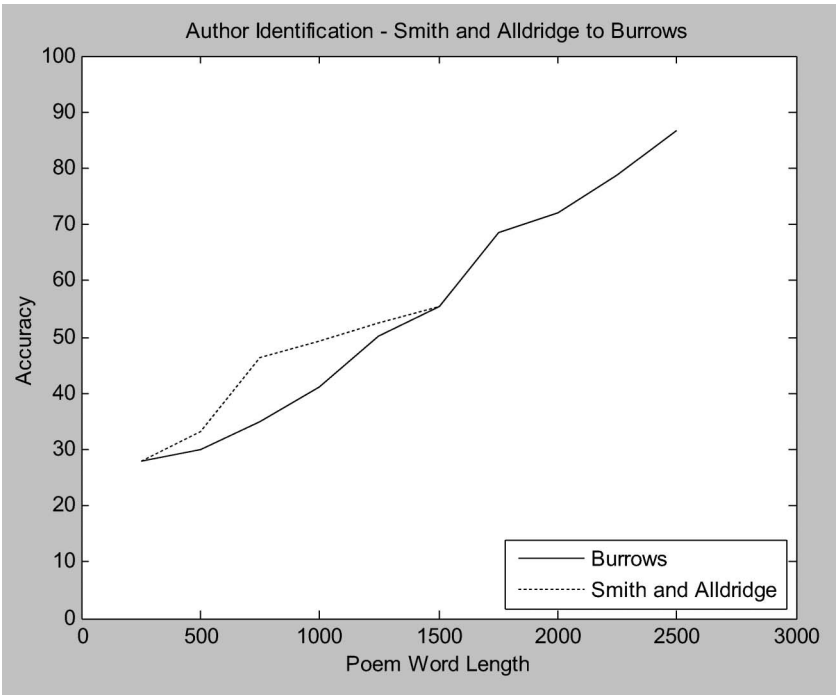

Fig. 5. Correct author identification compared with Burrows' results.

As is evident from Figures 7 and 8, it is clear that the accuracy of authorship identification increases with text length. The effectiveness of the method is under 80% for poems shorter than 2000 words. The selection of poems chosen for this experiment consisting of more than 2000 words is less than 5% which confirms Burrows' original observation that the method has limited use within this genre (Burrows, 2002). However, the results for author short-listing are more promising, with an accuracy of over 80% on poems of over 500 words. In the experiments, we determined that all results were statistically significant at the 5% level, with the exception of the improvements observed when comparing results on poems of length 500–1000 words and 1000–2000 words.

The Impact of Word Vector Length on Authorship Accuracy

In the next stage of our experiments, we compared the impact of increased word vector size against text length. The results clearly demonstrate that text length is a far more important factor in authorship

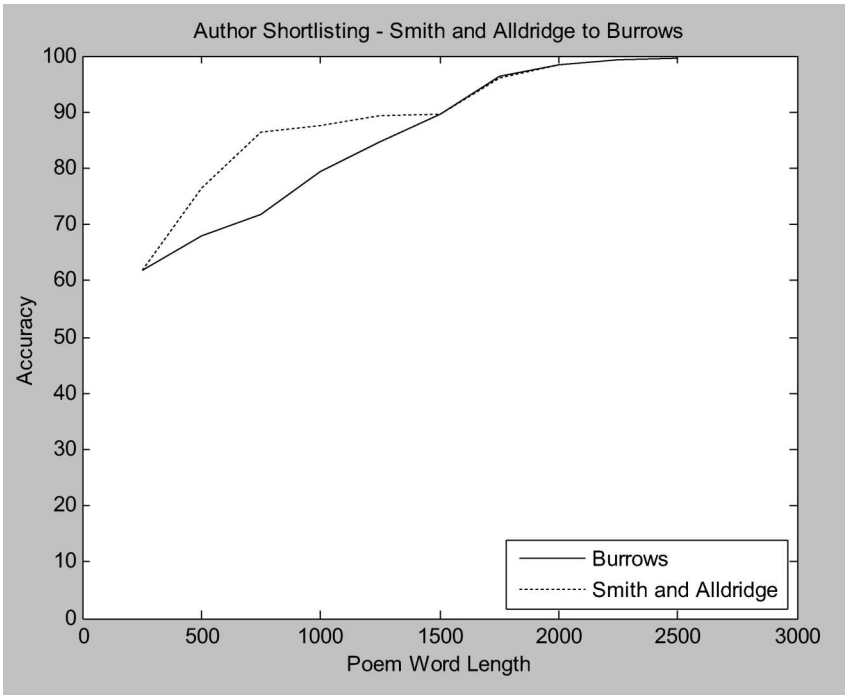

Fig. 6. Correct author short-listing compared with Burrows' results.

attribution than word vector length and are summarized in the ribbon graph given in Figure 9.

Figure 9 shows graphically that variance in word vector size is small compared with variance in text length accuracy. Figure 10 shows the variance in accuracy for texts of different length by varying word vector size.

Figure 10 also shows greater variance as the word vector length increases. Furthermore it shows an accuracy improvement for author identification when the word vector sizes are increased to 300. However, with word vector sizes above 300 words the accuracy of authorship identification becomes steadily worse as can be seen from median bar in the box-plot.

Overall, we can see that performance for the Delta score improves as word vector size increases to about 300, but after that becomes worse as it is increased up to 800. The differences in accuracy between word frequency vector length and text comparison size are summarized in the chart in Figure 11.

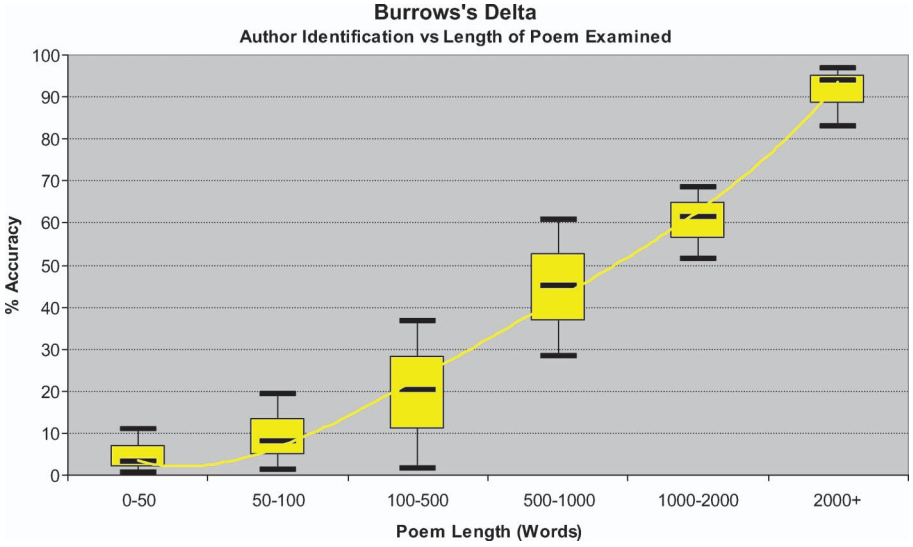

Fig. 7. The effect of poem length on authorship attribution accuracy.

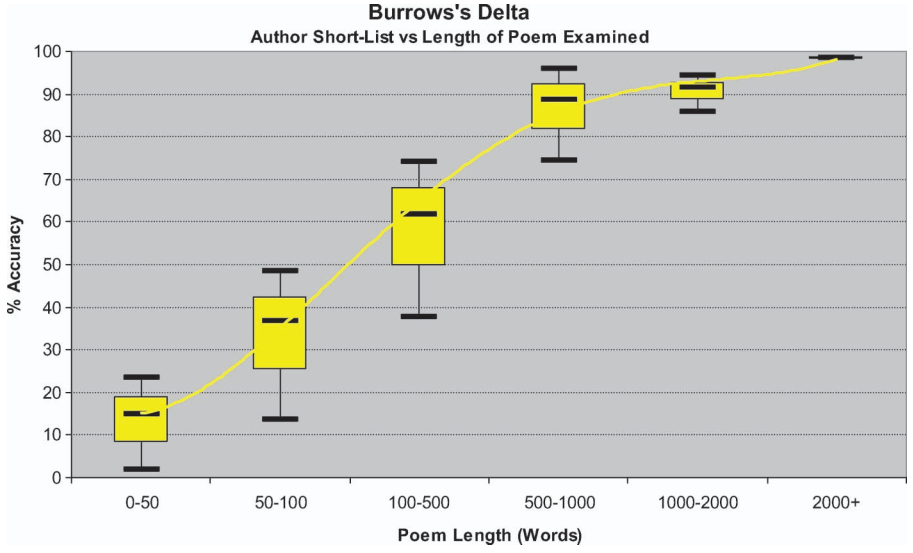

Fig. 8. The effect of poem length on author short-listing.

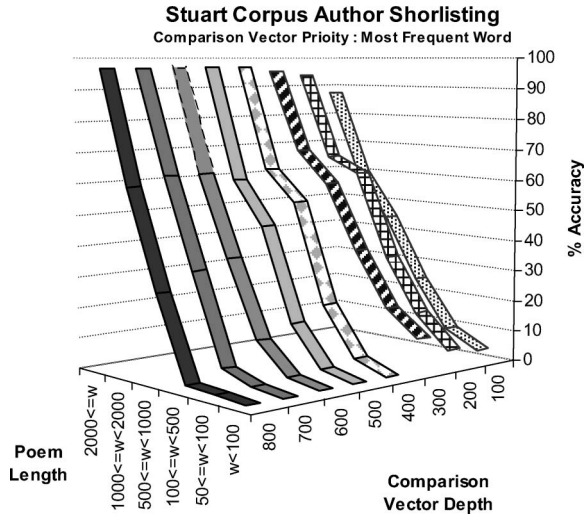

Fig. 9. Comparison of word length and word vector size.

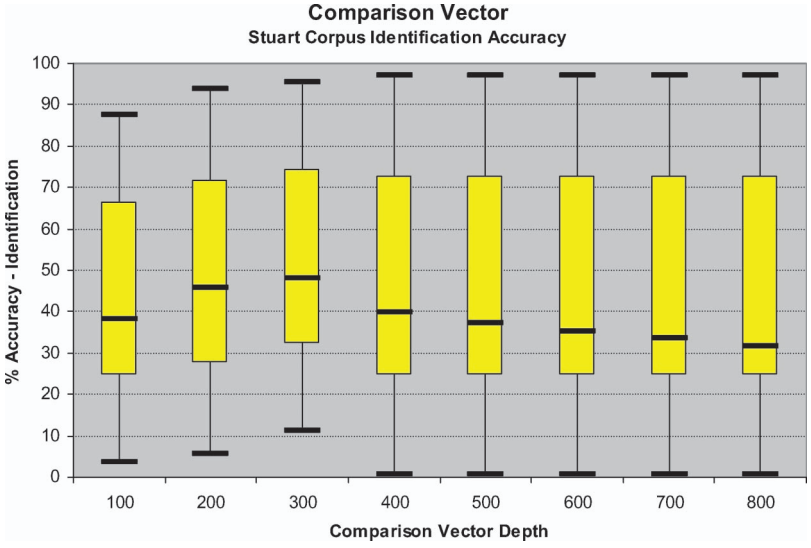

Fig. 10. Variation in accuracy for different word vector sizes.

Figure 11 clearly demonstrates that the lowest accuracies occurred with the largest word vectors and the smaller word vectors seem to give the largest overall accuracy. This observation applies to texts of all lengths, but is most marked when the text length is less than 500 words.

This differs from Hoover's finding when testing Burrows' Delta on prose (Hoover, 2004b). The Hoover experiment was undertaken on prose with longer texts, where the difference was not so marked. Our experiments indicate that the word frequency vector worsens when the text length under consideration contains fewer words than the number of word dimensions in the word frequency vector. This suggests that over-fitting might be occurring for larger word frequency vectors on short texts.

The Effect of Altering the Starting Point of the Word Vector

The word vector has generally been set at the *n* most frequently occurring words, without any real consideration as to whether that is the optimum

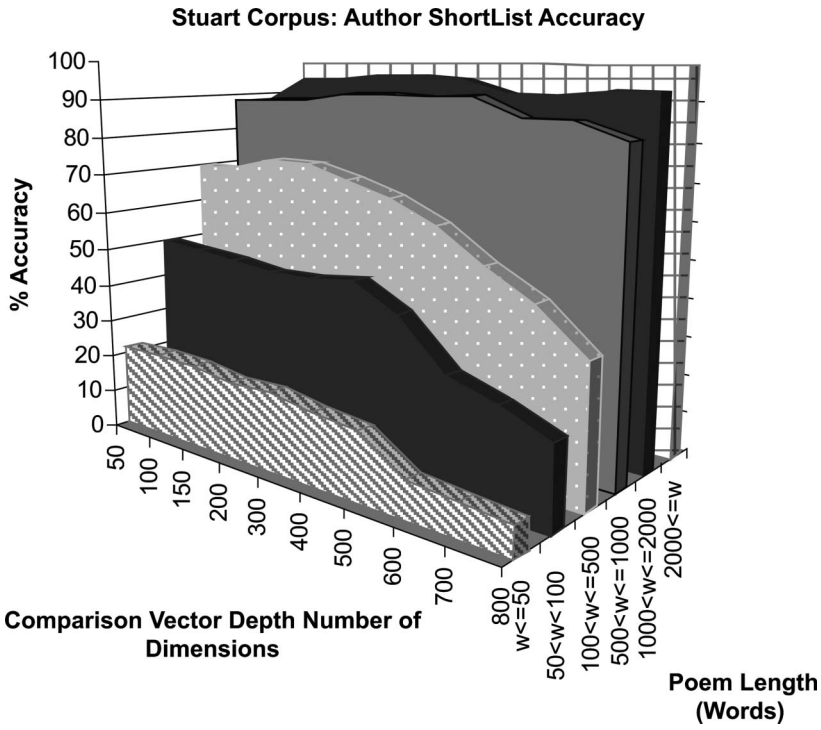

Fig. 11. Accuracy of word vector depth for different text lengths.

starting point. The aim of this experiment is to vary the starting point but using a fixed size word vector of 200 words. The results are shown in Figure 12. The values along the *x*-axis indicate a word vector comprising $n+200$ words, where n is the offset from the most frequent word and words 1 to $n-1$ are excluded. That is for $n=100$, the 100th to 300th most frequent words would be used in the word frequency vector and the 1st to 99th most frequent words would be excluded.

The offset of 1, represented by the leftmost box-plot, represents a word vector of size 200, comprising the 1st to 200th most frequent words, and an offset of 51 represents a word vector comprising the 51st to 250th most frequent words. There is a clear degradation of accuracy as the offset is increased, demonstrating that high frequency words are important in authorship identification. The experiment was repeated on differing word vector sizes (e.g. 50, 150 and 400 words) and we found consistently worsening performance as the offset was increased. Our experiment demonstrates that use of most frequent words starting at the highest frequency word is very important to the success of using Burrows' Delta.

Choice of Words in the Word Vector – Other Approaches

In order to further examine the selection of words for use in the word vector, we examined the effect of altering priorities for word choice to

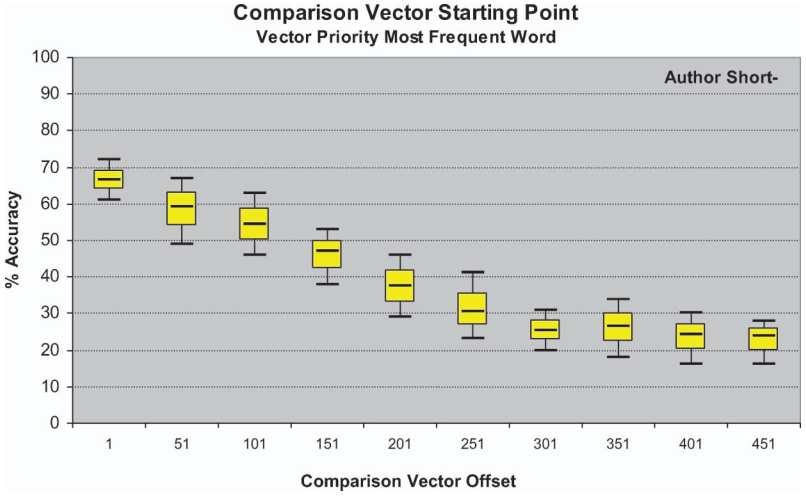

Fig. 12. Comparison by vector starting point.

analyse the effect on performance. The most frequent word priority used in this experiment is consistent with the method used by Burrows (2002), but other orderings are possible. Table 1 summarises our experiments.

Figure 13 clearly demonstrates that MFW is the ordering that provides the greatest accuracy. The word frequency vector listing function words first is interesting because although the performance is worse, the range is much less. Selective exclusions such as MFW-Odd Ordinal increase the error and significantly reduce accuracy. The word frequency vector for author short-listing gave similar results, although the degradation of performance for author short-listing was not as pronounced. The differences were significant at the 5% level with the exception of the difference between MFW and MFW-F/C. When we tested each word frequency vector type with texts of differing lengths the results were consistent with previous tests in that the most accurate results were obtained with the evaluation of poems over 2000 words in length and accuracy progressively deteriorated as the number of words in the poem

Table 1. Changing comparison vector priorities.

Ordering priority	Definition of word frequency vector
Most frequent word (MFW)	The top *n* most frequently occurring words in the corpus.
MFW – F/C	The top *n* words taken with the corpus ordered on most frequent word, but placing all function words before content words.
MFW Odd Ordinal	The top *n* words taken with the corpus ordered on most frequent word, selecting only odd ordinal positions (1st, 3rd, 5th, ...)
MFW – C/F	The top *n* words taken with the corpus ordered on most frequent word and placing all content words before function words.
Lexeme length	The top *n* words ordered on lexeme length, with short words first.
LFW – F/C	The top *n* words taken from the 1600 most frequent words after sorting on least frequent word and placing all function words before content words.
LFW – C/F	The top *n* words taken from the 1600 most frequent words after sorting on least frequent word and placing all content words before function words.
Random	The top *n* words taken from 1600 most frequently occurring words after ordering randomly.

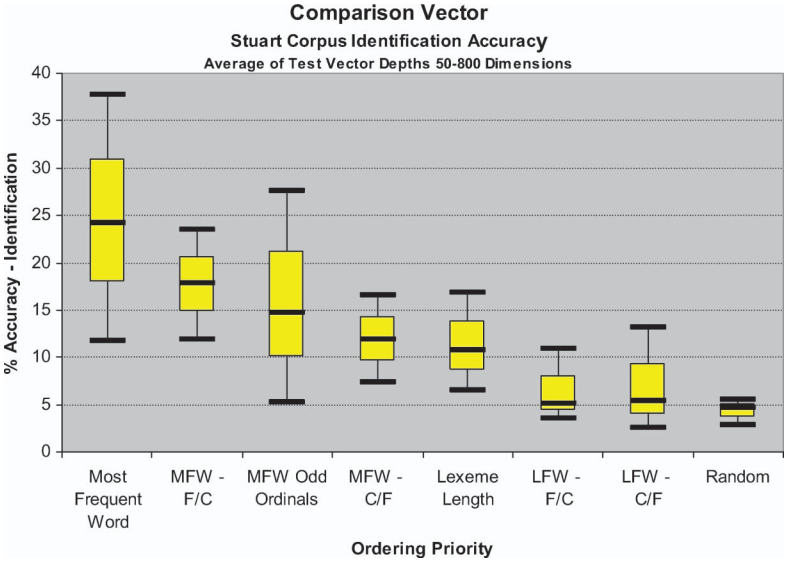

Fig. 13. Changed word vector priorities for authorship identification.

to be evaluated was reduced. The most accurate results were obtained with the MFW and MFW-F/C word vectors.

We compared word vector size with word vector priorities and the results are summarised in Figure 14.

The best performance is provided by MFW and MFW-F/C providing comparative performance at word vector sizes up to 150 and above 300 words. Between 150 and 300 words, MFW-C/F outperforms MFW-F/C suggesting that the most frequently occurring content words contribute to accuracy gains observed by Hoover (2004b), but the results are not statistically significant in our tests at the 5% level. It is noticeable that once again, performance drops off as word vector size is increased much beyond 300 words. Once again we suspect that this fall-off in accuracy may be due to over-fitting.

A Comparison of Similarity Measures

In the next part of our experiments, we test angular similarity. In this case, we use cosine similarity which is a measure of similarity between two vectors of n dimensions by finding the cosine of the angle between the vectors. This measure is frequently used in text mining applications;

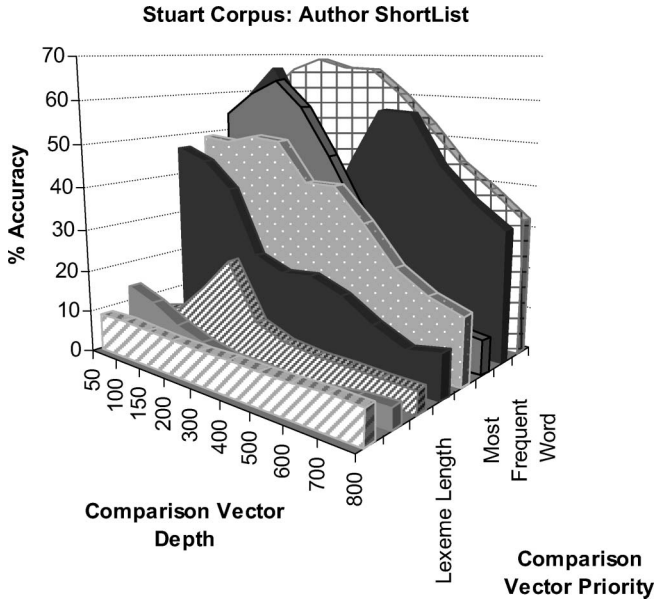

Fig. 14. Word vector size against word vector composition.

see Korfhage (1977) for more details on the cosine similarity measure. We compared it with the original Euclidean distance used by Burrows (2002) and in subsequent studies (e.g. Hoover, 2004a). Angular similarity compares the vectors in terms of their angular separation and ignores the scalar element of the vector which is expressed when Euclidean distance is calculated. The results are given in Figure 15.

Figure 15 demonstrates a clear improvement in both authorship attribution and short-listing, with differences being significant at the 5% level.

Figure 16 compares performance of the angular similarity measure and Euclidean distance over a range of different length poems and the difference in performance is evident and is particularly striking for short length texts. This is possibly not surprising as it is well known from experience in text mining that large word vectors show greater reliability using angular similarity measures (Frigui & Nasraoui, 2003). Figure 16 conclusively demonstrates that an angular similarity measure should be used in preference to Euclidean distance on Burrows' Delta.

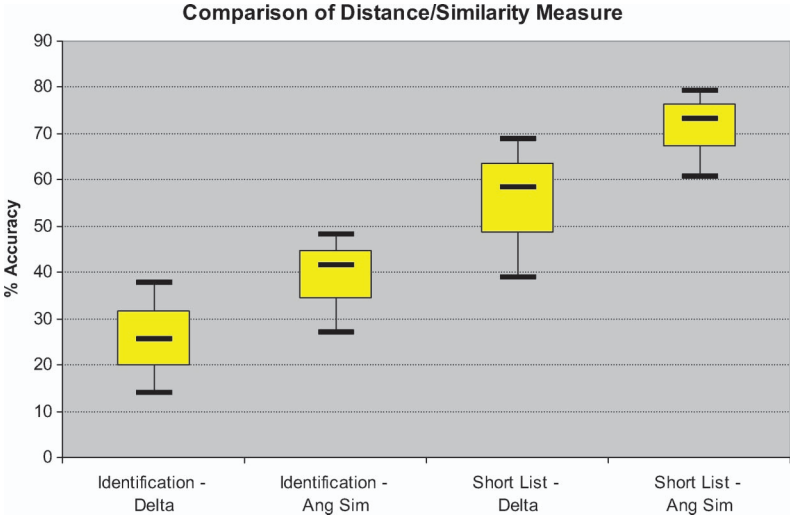

Fig. 15. Comparing angular similarity to Euclidean distance.

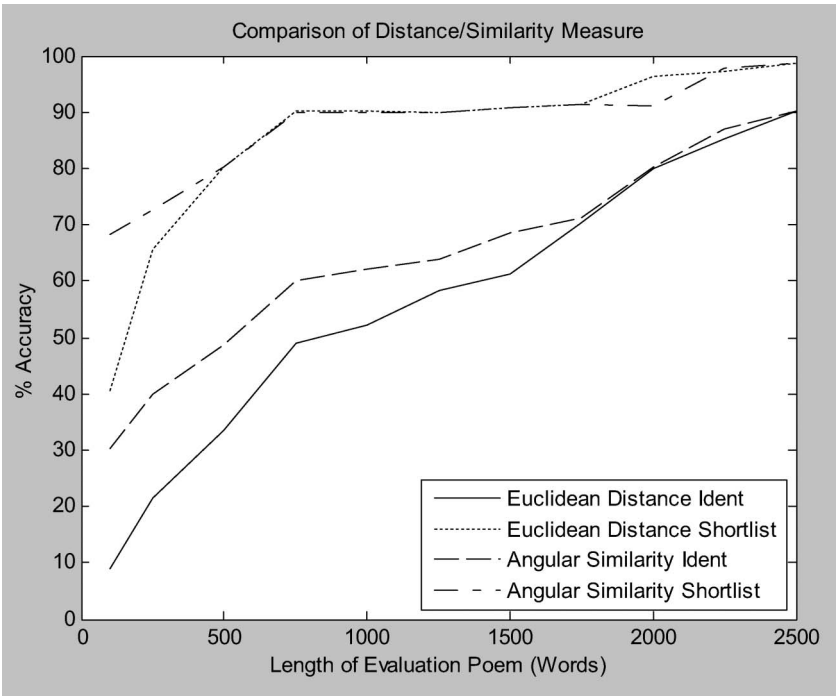

Fig. 16. A comparison of performance with angular similarity and Euclidean distance.

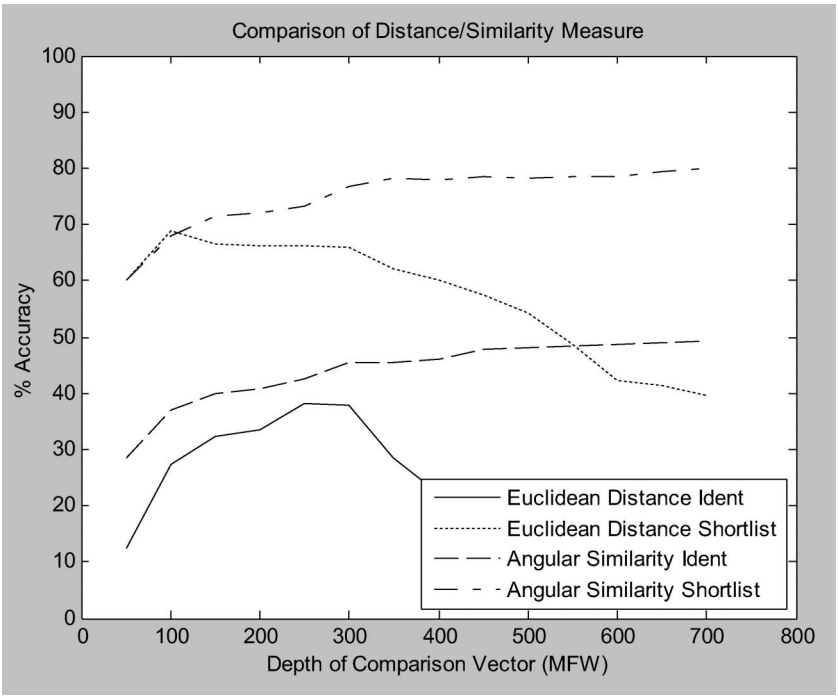

Fig. 17. A comparison of similarity measure over word vector sizes.

Similarity Measure and Word Depth Vector Comparisons

Angular similarity was universally more accurate on experiments where the word frequency vector depth was varied and the difference became very marked for larger word vector sizes. This is exactly as predicted from the known behaviour of Euclidean distance over higher dimensionality space and again confirms findings from text mining.

Figure 17 clearly demonstrates the superiority of angular similarity particularly over larger word vector sizes. We also compared word vector priorities using Euclidean and angular similarity measures. The results are shown in Figure 18.

The use of angular similarity, as well as outperforming Euclidean distance appears to show some resilience to differences in word frequency vector priorities. The result for random word vector choice is particularly striking.

The experiments conducted clearly demonstrate that an angular similarity measure should be used in preference to the Euclidean distance

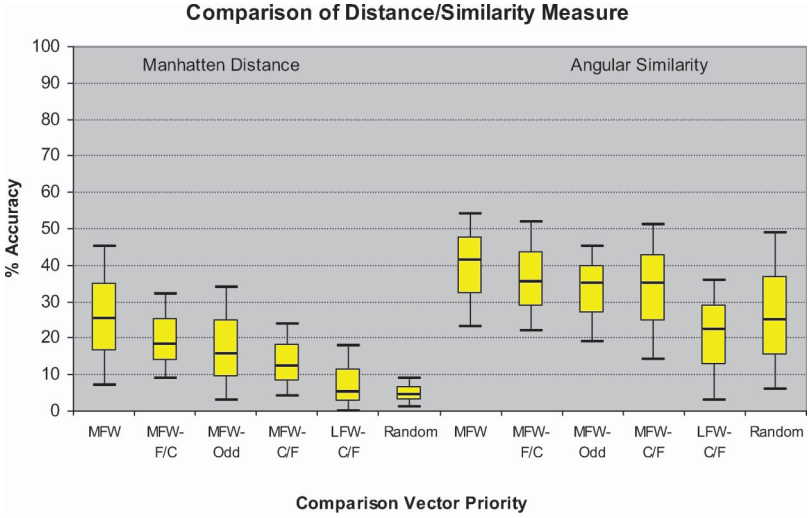

Fig. 18. Similarity measure and word vector depth priority.

measure that is routinely used. It also suggests that increasing the word vector beyond 200–300 words is likely to counter-productive.

Choice of the Most Appropriate Corpus

Bailey's Dictum states that authorship choice should be made by confining the study to a narrow set of authors and comparing them on a one-by-one basis (Bailey, 1979). On the other hand Burrows' Delta is designed to shortlist candidate authors from a list.

Our previous results have presented accuracy by comparing candidate texts with a corpus which is at most 50 years to either side of any author requiring identification or short-listing. We extend the corpus by including an additional 24 authors covering up to the 20th century. This corpus allows potential comparison with authors who lived more than 200 years after the restoration period. By conducting this experiment, we seek to establish how important it is to choose the correct corpus for use in the Delta method.

Figure 19 shows a marked deterioration of performance due to the extended corpus, with results that are significant at the 5% level. We found that accuracy still increases with increased text length, but it is compromised when the extended corpus is used, although proportionally more so on shorter texts. We also determined that accuracy is

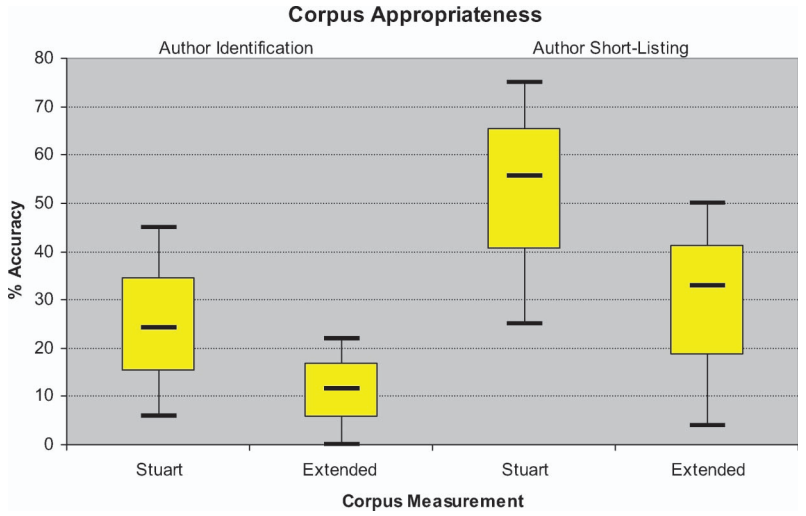

Fig. 19. Choice of corpus.

compromised regardless of word vector choice and regardless of similarity measure chosen.

WORD VECTOR CHOICE AND THE NON-AUTHOR PROBLEM

In a further experiment, we used principal component analysis (PCA) in order to further understand the role of words within the word vector. We carried out a PCA on the individual comparison scores for the 200 most frequent words for all poems within the Stuart corpus. The PCA identified six components that accounted for more than 99% of the variance. In fact using Kaiser's criterion we could have used a cut-off point of the first two components as they accounted for 90% of the variance. In the first component, which alone accounted for over 80% of the total variance, we noted that the second, seventh, eight and tenth deciles were contrasted from the others, suggesting that words in these deciles may be significant in author identification. Generally, we noted that the PCA loadings tended to emphasize the use of function words and de-emphasize the use of content words in the lower deciles. However, significant emphasis was placed on content words that appeared in the second decile.

Overall, it appears that function words contribute more to variance, but the importance of high-frequency content words, particularly those in the second decile should not be underestimated.

Regardless of whether angular similarity or Euclidean distance is used, one problem that the Delta method faces is the problem of mis-identification because a candidate author is absent from the author comparison list. We also decided to explore the problem of word vector choice further using a different method of authorship identification.

In the final experiment we explore the problem of word vector choice and the difficulty of authorship attribution when an author other than the ones chosen for comparison is responsible. In its simplest form, it can be treated as a two-way classification problem that also requires a non-classification result. Because the technique necessitated a larger volume of text we chose different authors for this study – Jane Austen, Charlotte Bronte and Emily Bronte – and used discriminant analysis for classification. The word vectors used were chosen in different ways:

- Removing inconsistent words across texts by the same author.
- Identifying consistent words across texts by the same author.
- Stepwise introduction of variables.

We chose the six major novels by Jane Austen (*Pride and Prejudice*, *Sense and Sensibility*, *Mansfield Park*, *Northanger Abbey*, *Persuasion* and *Emma*), and four major novels of Charlotte Bronte (*Jane Eyre*, *The Professor*, *Shirley* and *Villette*). We also used *Wuthering Heights* by Emily Bronte for unknown author classification. The novels of Jane Austen and Charlotte Bronte were used as a training set, either using the full set of novels or leaving out one novel. Three texts were used as a test set: *Lady Susan* – an early work by Jane Austen, *A Biographical Notice of Ellis and Acton Bell* by Charlotte Bronte and *Wuthering Heights* by Emily Bronte.

The method used was to create a discriminant function using a training set. The discriminant function was used to separate the writing of Charlotte Bronte and Jane Austen. Having created the discriminant function, an allocation rule was created by finding the cut-point of the discriminant function using the mean of the functions at group centroids. The discriminant function is then applied to the test data, the cut-point applied and the allocations tabulated.

The earlier experiments revealed that most frequent words appear to provide the most accurate classification. Although the choice of an optimum word vector set is ultimately a combinatorial optimisation problem. The Burrows' Delta method relies on the use of a corpus to generate a word vector for use in authorship identification. We suggest that the Delta method may be useful in situations where candidates for authorship are to be short-listed, but specific tests for individual authors should be used where possible. This raises the question of what word vector should be used for a given author.

Our approach involved creating an objective test to determine whether words should be included in the word vector. We start with the problem of creating a word vector for Jane Austen. The length of the novels and the volume of text involved enabled us to break the text into blocks and use word counts of each block. This generated a frequency distribution of words for each novel. The first criterion for inclusion was that each word that should be considered in the word vector appears in all novels. This allows an objective test for eliminating proper names and character names for individual novels. We then tried reducing the word vector set further by comparing word frequencies of novels paired together, for each word, treating the words as mean-adjusted data clusters and then measuring the intra-cluster distance for both clusters and comparing them against the inter-cluster distance. This allows us to describe the distribution as a ratio:

intra(cluster1) : intra(cluster2) : inter(cluster1,cluster2)

This ratio is a measure of consistency of use. We noted that for the creation of this word vector, word usage in *Pride and Prejudice* appeared inconsistent compared with Jane Austen's other novels. We initially allowed no more than a 10% variation in the ratios, which generated a small word vector for Jane Austen of only 14 words: {she, at, no, now, should, know, most, two, away, enough, till, least, certainly, supposed}.

However, when this word vector was used both to identify previously unseen works by Jane Austen and to separate Jane Austen from Charlotte Bronte, results were rather disappointing. We reverted back to the word vector that filtered out words that did not appear in all Austen novels. This word vector was then used as the basis for a series of experiments.

We used the stepwise introduction method of discriminant analysis to introduce variables one-by-one into the discriminant function using Wilk's Lambda as a criterion for selection because it takes into consideration both homogeneity within groups and distance between groups. However, this method is susceptible to local maxima and does not guarantee an optimum word vector choice.

We carried out the following experiments:

- Most frequent words.
- Most frequent function words.
- Most frequent non-function words.
- Most frequent pronouns.
- Most frequent verbs.
- Most frequent content words.

Figure 20 compares the performance of accuracy of discriminants to separate Jane Austen from Charlotte Bronte with the step introduction demonstrating greater efficiency, indeed using only a word vector of 61, 100% accuracy was achieved in discriminating between Austen and Charlotte Bronte. Using a word vector of function words only, 100% accuracy in separation was achieved using only the 38 top function words, although this word vector was then not as accurate when applied to the test set as the most frequent word vector of 61 was. Word vectors comprising non-function words or pronouns or content words produced far worse results.

Authorship Attribution as a Classification Problem

Authorship attribution is ultimately a classification problem and this raises several questions: firstly the question of whether multiple category classification should be used in preference to two-category classification. Burrows' Delta advocates a multiple classification, but it is by no means clear that this is preferable to a more simple two-way classification test in which a custom-made word vector is used for each author. In Smith and Rickards (2008) we use a genetic algorithm to optimise the word vector that achieves maximum separation between two authors – in this case, Thomas Jefferson and Thomas Paine – and when this optimised word vector is applied to the *Declaration of Independence*, it suggests that Thomas Paine is the author, not Thomas Jefferson.

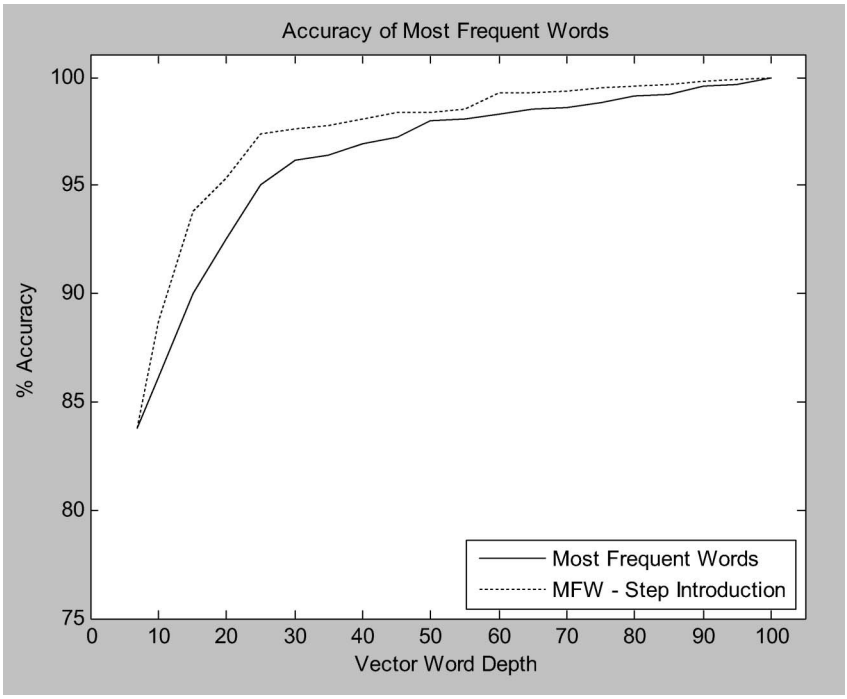

Fig. 20. Accuracy of most frequent words.

A difficult question for any authorship attribution study is the problem of non-classification. In our study, we created a discriminant function based on a custom-built word vector, created by examining homogeneity of usage for texts by an author with known provenance, however when we applied Emily Bronte's *Wuthering Heights* to a discriminant vector created to differentiate between Charlotte Bronte and Jane Austen, the resulting discriminant function applied to *Wuthering Heights* assigned authorship to Charlotte Bronte with 100% accuracy. In fact the resulting discriminant had a value that was "more Charlotte Bronte than Charlotte herself". This gives rise to the problem of non-classification, i.e. given two authors A and B and a text written by C, how do we achieve a result that will not classify the text as being by either A or B?

In Smith and Rickards (2008) we define an *n*-space object for each author using a word vector that maximises the separation between two authors and then enclose the *n*-space using an *n*-dimensional convex hull

which divides the *n*-space up into three regions: an enclosing space for author A, an enclosing space for author B and everywhere else in the *n*-space which is classed as non-authorship attribution. In this way we can also specify non-classification. We suggest that this work can be extended using a tailor-made single author word vector.

SUMMARY

We carried out a systematic analysis of Burrows' Delta method. Our findings were as follows. Firstly word vector choice by most frequent word leads to the greatest degree of accuracy and furthermore the word vector size is optimal for sizes of 200–300 words. Beyond 300 words, accuracy is degraded. Significant improvements in both author short-listing and authorship attribution can be achieved by using an angular similarity measure in place of the Euclidean distance that is customarily used. This suggests quite strongly that Burrows' Delta should be adopted to use the cosine similarity measure. The importance of correct choice of corpus for comparison is demonstrated and our findings suggest that function words contribute considerably to authorship attribution, but high frequency content words also appear to play a role. Further reductions in word vector size may be possible by optimisation of the word vector and further accuracy gains may be made by tailor-made word vectors for individual authors.

REFERENCES

Aldridge, W. (2007). *The Burrows Delta Dilemma: Optimization of Delta for Authorship Attribution*. MSc thesis, City University, London.

Bailey, R. W. (1979). Authorship attribution in a forensic setting. In D. E. Ager, F. E. Knowles & J. Smith (Eds), *Advances in Computer-Aided Literary and Linguistic Research. Proceedings of the 5th International Symposium on Computers in Literary and Linguistic Research* (pp. 1–15). Birmingham: John Goodman.

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, *17*(3), 267–287.

Frigui, H., & Nasraoui, O. (2003). Simultaneous clustering and dynamic keyword weighting for text documents. In M.W. Berry (Ed.), *Survey of Text Mining* (pp. 45–72). New York: Springer.

Gardner, H. (Ed.) (1972). *The New Oxford Book of English Verse*. Oxford: Oxford University Press.

Hoover, D. (2004a). Testing Burrows' Delta. *Literary and Linguistic Computing*, *19*(4), 453–475.

Hoover, D. (2004b). Delta prime? *Literary and Linguistic Computing*. *19*(4), 477–495.

Hoover, D. (2006). Word frequency, statistical stylistics and authorship attribution. Word frequency and keyword extraction. *AHRC ICT Methods Network Expert Systems Seminar on Linguistics*. Lancaster University.

Korfhage, R. R. (1977). *Information Storage and Retrieval*. New York: Wiley.

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

Oldpoetry. Retrieved June 2007, from www.oldpoetry.com.

Parbat, S. (2008). Implementing Burrows' Delta using a gamma distribution. Undergraduate project, Department of Computing, City University London.

Project Gutenberg. Retrieved April 2008, from www.gutenberg.org.

Smith, P. W. H., & Rickards, D. (2008). The authorship of the American Declaration of Independence. *AISB2008 Convention: Symposium on Style In Text: Creative Generation and Identification of Authorship*. Aberdeen.