Searching with Style: Authorship Attribution in Classic Literature

Ying Zhao **Justin Zobel**

School of Computer Science and Information Technology
RMIT University,
GPO Box 2476, Melbourne, Australia,
Email: yizhao,jz@cs.rmit.edu.au

Abstract

It is a truism of literature that certain authors have a highly recognizable style. The concept of style underlies the authorship attribution techniques that have been applied to tasks such as identifying which of several authors wrote a particular news article. In this paper, we explore whether the works of authors of classic literature can be correctly identified with either of two approaches to attribution, using a collection of 634 texts by 55 authors. Our results show that these methods can be highly accurate, with errors primarily for authors where it might be argued that style is lacking. And did Marlowe write the works of Shakespeare? Our preliminary evidence suggests not.

Keywords: Stylistics, authorship search, language modelling, document management

1 Persuasion

The notion of style is central to literature. The best-known authors of classic English novels and plays are renowned for having distinctive styles that make their works immediately recognizable. From the complexities of Henry James, the humour of Dickens, and the directness of Austen to the folksiness of Twain and the simplicity of London, a reader who is familiar with particular novelists can easily recognize their writing. Some authors are read as much for their style and writing as for what they have to say.

Style is not easy to define or identify. However, it is on the notion of style that the task of authorship attribution (AA) depends: that some element of an author's writing can be used as a reliable marker of their work. Given such markers, AA techniques can be used to verify whether a particular work is by a particular author, or to identify a likely author from amongst a set of candidates. Applications include forensics, plagiarism detection, and analysis of literature.

Current AA techniques have two components: an indexing mechanism for extracting style markers from text, and a comparison mechanism for using the markers to determine probable authorship. The style markers that have been used for this task have been relatively limited; in most work the markers have been distributions of text elements such as function words (Burrows 1987, Binongo 2003, Baayen et al. 2002, Juola & Baayen 2003, Holmes et al. 2001, Zhao & Zobel 2005), punctuation symbols (Baayen et al. 2002), and part-of-speech tags (Kukushkina et al. 2000, Stamatatos et al. 1999, 2001, Baayen et al. 1996, Zhao et al. 2006).

Comparison mechanisms have been more diverse. Most of the methods are based on statistical analysis, such as principle component analysis (PCA) (Baayen et al. 1996, Holmes et al. 2001, Burrows 2002), linear discriminant analysis (Baayen et al. 2002, Stamatatos et al. 2001); and machine learning techniques, such as Bayesian networks and support vector machines (SVMs) (Diederich et al. 2003, Koppel & Schler 2004), and treat AA as a classification problem. Some work is driven by particular AA problems that researchers have chosen to investigate, and most investigations have involved small volumes of data and small numbers of authors, such as the 65 Federalist Papers of known authorship, each written by one of two people (Fung 2003, Khmelev & Tweedie 2002).

In previous work we have investigated several aspects of AA, including style markers and comparison methods (Zhao & Zobel 2005, Zhao et al. 2006). We have found, in agreement with other researchers, that function words are a reliable indicator of authorship. Using newswire data, we have explored several AA methods, finding that the best results are yielded by SVMs and statistical methods based on language models and entropy (Zhao et al. 2006). Methods such as SVMs can be effective, but are not efficient; it is far from obvious that they can be scaled to the data collections found in typical text repositories. In work in progress, we have explored a search-based method of AA, in which language models are used to match documents by style rather by, as is the usual case in text search, content (Zhao & Zobel n.d.). This approach—also tested on newswire data—can be used for a collection of 500,000 documents. However, even though these methods are successful on average, they are not successful in some particular cases. Why this occurs has been unclear.

In this paper, to further explore the properties of AA methods, we apply them to a corpus of novels extracted from the Gutenberg project. While not a large corpus by text collection standards, it is more substantial than the collections used in most previous work for AA, and contains a substantial cross-section of 19th-century English literature as well as other work. Using this collection, we explore the use of three types of style markers—function words, part-of-speech tags (POS), and POS pairs—and the combination of these.

Our results show that authorship can be attributed with high reliability, with classification substantially outperforming search-based AA, while function words are more effective than other types of style markers. Overall results for the best method, using complete texts as queries, give attribution accuracy on positive examples of over 85% and on negative examples of over 95%. Use of parts of texts was less

10% overall accuracy, while classification on 10,000-word fragments achieved 53% accuracy—not as good as with complete documents, but sufficient to give an indication of likely authorship.

The errors, that is, cases where texts are misattributed, are illuminating. The commonest error is to misattribute a text that was not originally written in English, suggesting that style—as measured by our methods—does not survive the translation process. A difficult author was Wilde, a result that is perhaps unsurprising given that he was a satirist and thus an imitator of other people's styles. Another difficult author was Defoe, the earliest novelist in our collection. Other errors were not so easily classified, but an interesting (though weak) trend was to misattribute to another author from the same period. Overall, these errors show that the problems in AA probably lie as much in the work as in the method: when given texts that are expected to have identifiable style, AA appears to be highly reliable.

Revisiting a well-known AA problem, in our collection we included plays by several major playwrights of the late sixteenth and early seventeenth century: Marlowe, Jonson, Beaumont & Fletcher (who wrote together), and Shakespeare. In the rare cases that these works were misattributed by the methods we used, they were attributed to another of the group, not unsurprisingly given the changes in English between these works and those of the later authors that made up the bulk of our corpus. However, if Marlowe wrote Shakespeare, as has sometimes been speculated, there is little evidence for it here.

2 A Study in Stylistics

Authorship attribution (AA) is the process of attempting to identify the likely authorship of a given document, given a collection of documents whose authorship is known. Most of the methods described in the research literature consist of two components, an indexing mechanism and a comparison mechanism. The indexer converts each document to a set of tokens or markers whose properties are assumed to be characteristic in some way of a particular author. The comparator uses these markers to assign an author to unattributed documents.

Authorship attribution is an example of use of stylistic aspects of text in retrieval (Pol 2005, Sarkar et al. 2005, Kaster et al. 2005). Style concerns the way in which a document is written rather than its contents; stylistics is the study of style. Automated analysis of stylistics can be applied to a range of problems, from document attribution and authentication to matching document readability to the abilities of the user.

All published AA methods make use of collections of training data of known attribution. These are used to establish models of one kind or another. One data collection is the 65 Federalist papers, written during the debate that led to the creation of the US constitution. As another example of this kind, Holmes et al. (2001) used a collection of 17 journal articles by two authors. Other researchers have used data collections developed specifically for AA research. For example, Baayen et al. (2002) asked eight students to write a total of 72 articles. In our previous work (Zhao & Zobel 2005), we used attributed articles drawn from newswire data, focussing on authors who had made several hundred contributions each. Our motivation was to have a data collection of realistic difficulty both in size and kind. As authors of news articles aim primarily to communicate rather than create art, and as news articles are typically no more than a few thou- would be more significant than on some other data sets that have been used. In work in progress we report on successful AA on a collection of 500,000 newswire articles (Zhao & Zobel n.d.).

These datasets are used to test AA in different ways. Some AA problems are two-class, that is, the collection and the unknown documents are all by one of two known authors. Some AA problems are multi-class, which is the generalization of two-class to multiple authors. Some AA problems are one-class, in which some of the documents are by a given author and the rest are unknown; the task is to identify whether a new document is or is not by the given author.

Within computer science, the focus of research has been on comparators rather than indexers, with most researchers assuming a straightforward indexing method and using it as input to a comparator. A common indexing method is to extract function words (or closed-class words) (Burrows 1987, Baayen et al. 2002, Juola & Baayen 2003, Holmes et al. 2001, Kukushkina et al. 2000, Binongo 2003). An alternative is to use NLP methods to annotate the text with parts-of-speech (POS) tags (Baayen et al. 1996, Kukushkina et al. 2000, Stamatatos et al. 1999, 2001, Li et al. 2006, Masuyama & Nakagawa 2004), and to use these tags—or sequences of tags—as markers. Use of POS is intuitively attractive, but work to date has found POS tags to be no more effective than function words, a result that is confirmed in this paper.

Attribution is a form of classification, and thus it is attractive to apply existing classification methods; many of the proposed AA methods are based on classification techniques. Classification has been investigated in areas such as machine learning (Scholkopf & Smola 2002, Witten & Frank 2000, Quinlan 1993), text categorization (Bekkerman et al. 2003, Gabrilovich & Markovitch 2004, Khmelev & Teahan 2003, Li et al. 2003), and speech recognition. For AA, a range of classification-based attribution methods have been proposed. Binary authorship attribution is the simplest case. Binongo (2003) used PCA to investigate the writing pattern of the fifteenth book of Oz, as a case study of binary classification. In another case study, Holmes et al. (2001) also used PCA to distinguish between two authors. Multi-class and one-class AA are considered to be harder problems. Diederich et al. (2003) applied SVM for multi-classification AA and reported accuracy from 60% to 80%. Fung (2003) used SVM for feature selection on the Federalist papers. Koppel & Schler (2004) proposed an "unmask" approach for one-class AA, based on case-by-case selection of individual features for each author, and achieved around 80% accuracy. All words, not just function words, were considered.

Some methods have made use of the full text of documents (Diederich et al. 2003, Benedetto et al. 2002) rather than markers such as function words. Despite claimed good results, the plausibility of these methods is questionable, as they are based on word-occurrence statistics and choice of words is then as much a product of topic as of style; in one reported case an attempt to reproduce the results failed (Goodman 2002), and in unreported experiments we found that attribution based on full text of newswire articles was a complete failure.

A difficulty in examining much of this past work is that the methods were tested on different collections, making comparison of results far from straightforward. We compared a selection of these methods on our newswire collection (Zhao & Zobel 2005, Zhao et al. 2006) and found that Bayesian networks and SVMs were the most effective. However, there are challenges in scaling these methods to large volumes

has lower costs, as we now explain.

3 Priors and Prejudice

Classification methods such as those based on machine learning techniques make use of statistical properties of the items being classified. In computing, one of the most fundamental statistical properties is entropy. Intuitively, it seems plausible that the distribution of features in a new document should approximately match that of other documents by the same author. That is, we could build a model for each author based on known documents, and the model that is most like that of a new document can be assumed to identify its author.

On this basis, in previous work we have proposed an AA method using Kullback-Leibler Divergence (KLD)—a relative entropy measurement (Zhao et al. 2006). The underlying model of language is like that used in information retrieval (IR) (Lafferty & Zhai 2001, Zhai & Lafferty 2001, 2002, 2004). In this approach, the KLD between two models can be measured as:

$$KLD\left(p||q\right) = \Sigma_{x \in X} \;\; p\left(x\right) \log_2 \frac{p\left(x\right)}{q\left(x\right)} \qquad (1)$$

where $p\left(x\right)$ and $q\left(x\right)$ are probability mass functions used to calculate the probability of getting instance x. For AA, p is the model for new document d and q is a model based on known works of each author. AA involves computing $KLD(p||q)$ for p and every author model q, and choosing the q that gives the smallest relative entropy.

A simple form of model is the maximum likelihood $p(x) = f_{x,d}/|d|$, where $f_{x,d}$ is the number of occurrences of feature x in d and $|d|$ is the total number of feature occurrences in d. However, this has drawbacks in principle and in practice. Regarding these models as generative, the features that are observed in a document are a subset of those that might have occurred, and their absence from a document leads to obvious computational difficulties. Also, there is no control for the frequency of each feature in language in general. For example, if the features are function words, we would expect occurrences of a word such as "whilst" to be moderately indicative of authorship, while occurrences of a word such as "the" are uninformative. For this reason we need to use smoothing (Chen & Goodman 1996, Hiemstra 2002, Zhai & Lafferty 2001, 2004).

In our work, the probability mass function is formulated with Dirichlet smoothing, which is to date one of the most effective smoothing techniques in IR (Zhai & Lafferty 2004):

$$p'_d(x) = \frac{|d|}{\mu + |d|} p(x) + \frac{\mu}{\mu + |d|} p_B\left(x\right) \qquad (2)$$

Here, $p_B\left(x\right)$ is the probability of component x in a *background model*, which is used to address the problem caused by zero probabilities and to provide global feature statistics. The background model can be viewed as a collection of prior probabilities that can be used to bias the KLD to favour less common features. The parameter μ adjusts the significance of documents and background model contributing to the term weights. This parameter must be tuned; if it is too low, there is little discrimination between features, but if it is too high, the statistical properties of the document may be obscured. The value of μ can be critical for short documents where the statistical evidence is noisy; for long documents, which exact value of μ is relatively unimportant. The background model should ideally be drawn from a large collection of independent text.

Combining Equations 1 and 2, the dissimilarity $KLD\left(p_d||p_q\right)$ between two sets of features d and q can be written as:

$$\sum_{x \in q \cup d} \left[\left(\frac{f_{x,d}}{\mu + |d|} + \frac{\mu}{\mu + |d|} p_B(x) \right) \times \log \frac{\frac{f_{x,d}}{\mu+|d|} + \frac{\mu}{\mu+|d|} p_B(x)}{\frac{f_{x,q}}{\mu+|q|} + \frac{\mu}{\mu+|q|} p_B(x)} \right]$$

In the context of AA, we argue, use of KLD provides a principled approach where classification, rather than being based on elaborate statistical methods, is directly derived from simple fundamental theory (Zhao et al. 2006). Our results show that KLD is as effective as the best competitor method, SVMs, for binary AA, and is also effective for multi-class AA, a task for which SVMs are unsuited. In contrast to SVMs, the training cost is small, consisting only of counting of features.

Moreover, KLD-based AA can plausibly be applied to large collections. In work in progress (Zhao & Zobel n.d.), we have further explored relative entropy, but instead of finding the model that is most similar we use KLD as a way of ranking the documents in a collection according to the similarity of their feature sets to that of a query. This is, in principle, little different to standard text retrieval with a search engine (Zobel & Moffat 2006), and provides several potential advantages. The heuristics used during standard search can be applied, allowing rapid identification of the most similar documents (or rather, sets of features that represent documents) extremely fast. If the top-ranked documents are consistently by a given author, there is a clear indication of authorship of the query document; if, however, the highly-ranked documents are mixed, then it is likely that attribution is uncertainty. (An estimate of certainty is a useful guide to the quality of AA results.) Alternatively, given a query of known authorship, the matches are the documents most likely to be by the same author, an approach that has promise in for example plagiarism detection.

In work in progress (Zhao & Zobel n.d.), we have found that the KLD ranking approach is effective for authorship search. The largest collection we are using contains 500,000 documents, in which only 100 are positive examples by an author. The baseline probability of finding a single correct match in the top 10 is only 0.2%. We obtained an average of up to 44% of matches being correct in the top 10 documents in the ranking. It is therefore interesting to explore whether search works as an attribution method on other collections, as in the experiments described below. It is also interesting to explore how search compares to classification.

In this paper, our focus is on the problem of attribution in English literature. Are simple style markers sufficient for accurate identification of the author of a work? When attribution fails, what is the cause? Our experiments, described below, explore these questions.

4 Tests of the Diverse Styles

Our aims in this paper are to compare the effectiveness of search and classification as attribution methods; to see how effective attribution is on literature; and to understand when and why attribution fails. Given that we plan to use KLD for attribution, the

dexing method and a testbed dataset. We describe these below, then report on our experiments.

Sets and Sensitivity

A key aspect of this investigation was to explore the effectiveness of attribution on literature. The task may be relatively easy if the collection is small or there are only a few authors, or if the authors are from widely different periods. We sought to collect literature that was representative and consistent. Using the Gutenberg collection,[1] we gathered books from about 50 of the top-100 most downloaded authors. In most cases we collected 10 books, or fewer if less than 10 were available, but in some cases collected all works. The total number of books collected was 634, and the total number of authors (including playwrights as discussed below) was 55. We call this collection Gutenberg634.

In selecting the books, we avoided choices that we felt were inconsistent with the aims of our experiments. We did not collect volumes of poetry, dictionaries, or text in languages other than English. Individual short stories were avoided, especially in cases where a collection containing the story was also available. Authors with four or fewer works were not considered.

However, we did keep both plays and novels; plays were greatly in the minority but allowed us to examine attribution of the works of playwrights from the time of Shakespeare. The complete list of authors is shown in Table 2.

Maintenance of consistency was not straightforward. One book may have many different editions, or may be presented in different forms, such as both a complete edition and as a series of parts. The raw documents contain other-author material such as the Gutenberg disclaimer, editors' commentaries, and sometimes an introduction and preface written by someone else. We ensured in most cases that the major works of the author are included and that no duplicates are included. Finally, we manually deleted all other-author text from each book.

In the experiments, we built 634 collections of 633 documents each; that is, in each case one of the documents was left out to be used as a query (in other words, to be classified). In attribution, we can simply count the classification accuracy, based on N_c, how many documents were correctly classified. In search, other measures are possible. We use N_r, the number of documents by the same author ranked in the top 5. Given the maximum possible value for N_c (respectively, N_r), we can then give a percentage accuracy for each technique. The number of correct documents in the top 5, or precision at 5 documents returned, we denote as P@5. As can be seen in the tables, results are broken down by author and by method.

Measure for Measure

Indexing was one of the aspects of AA explored in our experiments. We tested several different simple forms of marker types. One form of marker was function words, which are content-free but grammatically important words such as prepositions, conjunctions, articles, and elements such as words describing quantities. Function words have shown to be reasonably effective for binary and multi-class AA in our previous work (Zhao & Zobel 2005, Zhao et al. 2006, Zhao & Zobel n.d.). We used a list of 363 function words.

Another form of marker was POS tags. We tagged the entire Gutenberg634 collection using NLTK (a natural language toolkit).[2] We used all tagged works of fiction from the Brown corpus[3] to train a unigram tagger. The accuracy of the trained tagger is 86.73%. In addition, we trained a Bigram tagger, but accuracy was only 83.23%, and we used the unigram tagger in our experiment. A list of 183 POS distinct tags was used. For example, for the text

> The widow she cried over me, and called me a poor lost lamb, and she called me a lot of other names, too, but she never meant no harm by it.

the function words extracted were "the over and a and a of other too but never no by it". The POS tags were "at nn pps vbd in ppo cc vbd ppo at jj vbn zz cc pps vbd ppo at nn in ap nns ql cc pps rb vbd at nn in ppo", in which, for example, nn is a noun.

We used the POS tags both individually and as pairs. The pairs should in principle give an indication of the way in which the author combines parts of speech, which is plausibly a signature of the author's style. However, automatic identification of POS may to some extent undermine this aim, as POS tagging is most likely to fail when presented with atypical word sequences and sentence formations—the very aspects of text that characterize style. That is, POS tags are likely to be least reliable for the most interesting elements in the text. In our results we also show the effect of combining the evidence of the different marker types, and thus have four sets of results in each of the two main experiments.

Table 1: *Usage statistics for the commonly used style markers for two authors. Each number is, for that author, the percentage of function word occurrences that is the particular function word. Counts are averaged across all documents available by each author.*

	Function words			POS tags		
	the	of	a	cc	in	jj
Shakespeare	7.6	4.8	4.1	3.8	5.9	2.8
Marlowe	9.5	6.2	3.2	3.2	6.4	2.4

Table 1 gives an example of how usage of different type of style markers can vary between authors. In this example from the collection of Gutenberg634, for even common style markers, the usage can be quite different.

As noted earlier, the smoothing parameter μ in Equation 2 plays an important role. We observed that, setting $\mu = 1000\sqrt{10}$ has given the best performance in our previous work (Zhao et al. 2006), in which we test binary AA with a small dataset of chapters of novels derived from the Gutenberg collection. (Chapters were extracted manually, an infeasible process with the larger Gutenberg634.) We used this value for parameter μ in our experiments here.

Choice of background model is another important factor in such experiments. Based on our experience with KLD in IR and AA, we believe that 634 books are not sufficient for a good background model. Therefore, we collected the background model from the AP collection, of over 250,000 newswire articles; AP is a sub-collection of the TREC data.[4] In some respects this choice is not ideal, consisting as it does of non-fiction written in the late 1980s and early 1990s, but was the best option available to us. This text

[1] See `www.gutenberg.org`.

[2] Available from `nltk.sourceforge.net/index.html`.

[3] The Brown corpus consists of one million English words gathered in 1961. The texts for the corpus are grouped into fifteen text categories. The corpus is the first of the modern computer readable corpora.

[4] See `trec.nist.gov`.

ments (Zhao & Zobel n.d.). As the results show, AA is highly effective with this background model; a better background model may further improve results, but they are already strong.

Great Expectations

Attribution via search provides a way of identifying the authorship of a new document, and of finding other documents that the author has written. Our aim in the first experiment was to test the effectiveness of different style markers when used in search-based AA. That is, the set of style markers extracted from each book is used as a query, and the other books are ranked according to their similarity. The hypothesis is that, if the markers are a good indication of style, then the highest documents in the ranking should be by the same author.

Each book was indexed in three ways: by function words, by POS tags, and by POS-tag pairs. For each form of indexing, we had 634 runs; in each, one of the books was used as a query and the remainder of the Gutenberg634 collection as a corpus. For each book, the rankings from the different forms of indexing were combined to give a fourth set of results. Performance for each query was measured with P@5, that is, the number of works by the same author in the top five ranked results. Outcomes, by author, are shown in Table 2. The "optimal retrieval" column shows the maximum number of correct results that could be obtained for each author. For example, Curtis wrote 7 books; the optimal result was 35; using function words, only 19 (or 54.3%) were retrieved; while results for Curtis using other style markers were somewhat lower.

As these results show, KLD-based search on markers is an effective mechanism for matching texts by authorship. Using function words as markers, on average over 76% of the documents in the top 5 are a correct match, and for 15 of the 55 authors accuracy is 90% or better. Other markers are somewhat less successful, but still reasonably effective, with 62% for POS tags and 66.5% for POS-tag pairs. Combination does no better than the average of the techniques being combined, at 71%. We had hoped that POS tags would prove the more effective method; and hypothesize, following the earlier discussion, that the very qualities that make an author's style unique may lead to tagging failures.

However, search-based AA is not particularly successful for some authors. An elementary cause might be the number of training examples; results were somewhat better for authors with more texts.

Other causes are attributable to style. Consider Schiller (German) and Tolstoy (Russian), two of the four authors whose works were originally written in a language other than English; the other two are Maupassant and Verne, both of whom originally wrote in French. Schiller and Tolstoy are amongst the worst cases for search-based AA with function words. Presumably the process of translation, or the fact that multiple translators may be involved, obliviates some of the individuality of style.

An interesting element in the errors is that there was a weak tendency for mismatches to be in the right period. For example, as discussed further below, when the works of Marlowe were used as queries, most of the matches were plays written by his peers.

Finally, to state the somewhat obvious, some authors do not have a strong writing style that can be easily identified easily, and other authors change their style between books. It is perhaps not surprising that the works of Wilde and Bierce, both satirists, prove difficult to attribute.

assessed the quality of attribution using the following rules. Given a query text by some author A, if three or more of the top 5 matches were by some author A', then we attributed the query text to A' with high confidence (and were right if $A = A'$). If two of the top 5 were by A' and the remainder were by three different authors, then we attributed the query text to A' with low confidence. Otherwise we judged the attribution to be unknown.

For function words, we attributed with strong confidence correctly in 451 cases and incorrectly in 61 cases, an accuracy of 88%. We attributed with low confidence correctly in 20 cases and incorrectly in 23 cases, not a wonderful result but much better that random. Attribution was unknown in 79 cases. Overall accuracy was 74%. It is against this result that a classification-based attribution method should be compared.

Through The Looking-Class

In our next experiment, we used KLD for one-class AA, on the same sets of markers. We again had 634 runs for each kind of markers. In each run, the aim was to make a decision on authorship— whether the query text was by a given author or was more likely to be by someone else. For example, for Austen we created a positive model using seven of her texts, created a negative model using the 626 texts by other authors, and used Austen's remaining text as a query. This was repeated for each of Austen's eight texts, and then for every other author. That is, classification was on a positive leave-one-out approach.

Results are shown in Table 3. As can be seen, classification is rather more effective than search-based attribution, achieving, for function words, an average of over 85%, in contrast to 74% for search-based attribution. Better than 90% accuracy is observed for 30 of the 55 authors. The POS-tag methods have proved much more successful than previously, but effectiveness is still slightly lower than with function words. Combination yielded little benefit. Improvement with POS tags requires, we believe, a more robust tagging method.

Similar failures can be observed. Schiller and Tolstoy are again problematic, as is Wilde. Another difficult author is Defoe, perhaps surprisingly, as he is the only author from the early eighteenth century. Overall, however, the results are highly satisfying.

Positive classification results give an estimate of the rate of false misses. Negative classification results are required to estimate the rate of false matches. That is, for a model trained on some author, say Austen, and a work by some other author, say Alcott, we wish to know the likelihood that the work will be attributed as by Austen. In these experiments, correctness is 95.2%, much higher than the accuracy on positive examples.

The texts we have used in these experiments are complete books, averaging over 80,000 words each. A question then is whether AA would be accurate on smaller texts. We re-ran the positive leave-one-out experiments using the full texts for training and, for each text being tested, a single 1000-word fragment as a query. Each fragment was drawn from a few thousand words after the start of the text. These experiments were not successful, with overall accuracy of only 10.4%. Use of 10,000-word fragments was more effective, giving overall accuracy of 53.2%. This result is far from perfect, but is much better than random, where average accuracy of around 3% would be expected. At this level accuracy, AA is not conclusive, but is nonetheless highly indicative.

Table 2: Results of authorship search experiments. Function words, POS tags, POS pairs, and combined features are used as style markers. Results are total P@5 per author and a percentage of optimal retrieval.

# of books per author	Optimal retrieval	Function words		POS tags		POS pairs		Mixed	
		N_r	$P@5$	N_r	$P@5$	N_r	$P@5$	N_r	$P@5$
Alcott(10)	50	43	86.0	32	64.0	32	64.0	38	76.0
Alger(10)	50	50	100.0	47	94.0	50	100.0	50	100.0
Austen(8)	40	39	97.5	31	77.5	37	92.5	38	95.0
Baum(10)	50	45	90.0	42	84.0	44	88.0	44	88.0
Bierce(8)	40	6	15.0	4	10.0	5	12.5	4	10.0
Burroughs(9)	45	39	86.7	21	46.7	27	60.0	33	73.3
Carroll(6)	30	7	23.3	4	13.3	1	3.3	3	10.0
Churchill(22)	110	93	84.5	75	68.2	78	70.9	87	79.1
Collins(23)	115	105	91.3	94	81.7	101	87.8	103	89.6
Conrad(12)	60	51	85.0	24	40.0	32	53.3	39	65.0
Curtis(7)	35	19	54.3	9	25.7	12	34.3	14	40.0
Darwin(9)	45	28	62.2	31	68.9	29	64.4	31	68.9
Defoe(9)	45	22	48.9	20	44.4	19	42.2	21	46.7
Dickens(11)	55	40	72.7	11	20.0	16	29.1	16	29.1
Fletcher(6)	30	23	76.7	19	63.3	20	66.7	22	73.3
Galsworthy(10)	50	22	44.0	27	54.0	29	58.0	30	60.0
Haggard(37)	185	168	90.8	154	83.2	165	89.2	168	90.8
Hardy(7)	35	35	100.0	18	51.4	15	42.9	18	51.4
Harte(9)	45	36	80.0	36	80.0	37	82.2	41	91.1
Hawthorne(10)	50	30	60.0	31	62.0	32	64.0	37	74.0
Henry(9)	45	40	88.9	37	82.2	40	88.9	40	88.9
Holmes(9)	45	30	66.7	20	44.4	20	44.4	25	55.6
Howells(10)	50	23	46.0	15	30.0	20	40.0	23	46.0
James(19)	95	80	84.2	48	50.5	44	46.3	58	61.1
Jonson(7)	35	19	54.3	26	74.3	30	85.7	29	82.9
Kingsley(10)	50	28	56.0	17	34.0	13	26.0	20	40.0
Kipling(8)	40	28	70.0	12	30.0	19	47.5	19	47.5
Lang(10)	50	19	38.0	11	22.0	14	28.0	14	28.0
Lever(9)	45	40	88.9	40	88.9	38	84.4	40	88.9
London(21)	105	101	96.2	64	61.0	79	75.2	93	88.6
Lytton(10)	50	49	98.0	49	98.0	43	86.0	49	98.0
MacDonald(9)	45	26	57.8	11	24.4	18	40.0	19	42.2
Marlowe(5)	20	6	35.0	6	30.0	8	40.0	9	45.0
Maupassant(9)	45	40	88.9	33	73.3	37	82.2	36	80.0
McCutcheon(10)	50	45	90.0	35	70.0	45	90.0	50	100.0
Motley(10)	50	50	100.0	50	100.0	45	90.0	50	100.0
Parker(10)	50	40	80.0	33	66.0	21	42.0	34	68.0
Pepy(10)	50	50	100.0	50	100.0	50	100.0	50	100.0
Poe(6)	30	21	70.0	18	60.0	19	63.3	22	73.3
Rohmer(10)	50	50	100.0	46	92.0	48	96.0	50	100.0
Schiller(10)	50	19	38.0	21	42.0	22	44.0	21	42.0
Scott(10)	50	50	100.0	49	98.0	50	100.0	50	100.0
Shakespeare(42)	210	203	96.7	197	93.8	199	94.8	201	95.7
Shaw(10)	50	33	66.0	28	56.0	30	60.0	30	60.0
Stevenson(10)	50	11	22.0	4	8.0	9	18.0	8	16.0
Stockton(10)	50	38	76.0	23	46.0	33	66.0	33	66.0
Tolstoy(15)	75	38	50.7	26	34.7	28	37.3	30	40.0
Twain(14)	70	57	81.4	33	47.1	46	65.7	50	71.4
Verne(10)	50	41	82.0	35	70.0	46	92.0	45	90.0
Wake(9)	45	34	75.6	40	88.9	38	84.4	40	88.9
Warner(10)	50	27	54.0	26	52.0	28	56.0	29	58.0
Wells(10)	50	23	46.0	17	34.0	23	46.0	22	44.0
Wilde(7)	35	2	5.7	2	5.7	1	2.9	2	5.7
Wodehouse(23)	115	113	98.3	97	84.3	100	87.0	102	88.7
Yonge(10)	50	33	66.0	13	26.0	21	42.0	21	42.0
Total(634)	3165	2409	76.1	1962	62.0	2106	66.5	2251	71.1

Table 3: Results of one-class authorship attribution. Function words, POS tags, POS pairs, and combined features are used as style markers. Results are total correct per author and a percentage of correct attribution.

Author	# of book	Function Words		POS tags		POS pairs		Mixed	
Name	in total	N_c	Accuracy	N_c	Accuracy	N_c	Accuracy	N_c	Accuracy
Alcott	10	9	90.0	9	90.0	8	80.0	9	90.0
Alger	10	10	100.0	10	100.0	10	100.0	10	100.0
Austen	8	8	100.0	7	87.5	7	87.5	7	87.5
Baum	10	10	100.0	9	90.0	9	90.0	9	90.0
Bierce	8	6	75.0	6	75.0	5	62.5	5	62.5
Burroughs	9	8	88.9	8	88.9	8	88.9	8	88.9
Carroll	6	3	50.0	3	50.0	2	33.3	2	33.3
Churchill	22	20	90.9	19	86.4	18	81.8	18	81.8
Collins	23	21	91.3	18	78.3	19	82.6	19	82.6
Conrad	12	12	100.0	11	91.7	11	91.7	12	100.0
Curtis	7	6	85.7	5	71.4	5	71.4	5	71.4
Darwin	9	6	66.7	6	66.7	7	77.8	7	77.8
Defoe	9	5	55.6	5	55.6	5	55.6	5	55.6
Dickens	11	8	72.7	6	54.5	6	54.5	6	54.5
Fletcher	6	6	100.0	6	100.0	6	100.0	6	100.0
Galsworthy	10	8	80.0	5	50.0	5	50.0	5	50.0
Haggard	37	26	70.3	31	83.8	30	81.1	30	81.1
Hardy	7	7	100.0	3	42.9	6	85.7	6	85.7
Harte	9	8	88.9	9	100.0	9	100.0	9	100.0
Hawthorne	10	5	50.0	9	90.0	8	80.0	8	80.0
Henry	9	9	100.0	9	100.0	9	100.0	9	100.0
Holmes	9	9	100.0	9	100.0	8	88.9	9	100.0
Howells	10	6	60.0	6	60.0	6	60.0	6	60.0
James	19	17	89.5	17	89.5	17	89.5	17	89.5
Jonson	7	7	100.0	7	100.0	7	100.0	7	100.0
Kingsley	10	9	90.0	9	90.0	9	90.0	9	90.0
Kipling	8	8	100.0	7	87.5	7	87.5	7	87.5
Lang	10	7	70.0	2	20.0	4	40.0	4	40.0
Lever	9	8	88.9	4	44.4	8	88.9	8	88.9
London	21	21	100.0	21	100.0	20	95.2	20	95.2
Lytton	10	9	90.0	10	100.0	10	100.0	10	100.0
MacDonald	9	7	77.8	5	55.6	7	77.8	6	66.7
Marlowe	5	5	100.0	5	100.0	5	100.0	5	100.0
Maupassant	9	7	77.8	8	88.9	7	77.8	8	88.9
McCutcheon	10	10	100.0	10	100.0	10	100.0	10	100.0
Motley	10	10	100.0	10	100.0	10	100.0	10	100.0
Parker	10	8	80.0	10	100.0	8	80.0	8	80.0
Pepy	10	10	100.0	10	100.0	10	100.0	10	100.0
Poe	6	6	100.0	6	100.0	6	100.0	6	100.0
Rohmer	10	10	100.0	10	100.0	10	100.0	10	100.0
Schiller	10	7	70.0	9	90.0	9	90.0	8	80.0
Scott	10	10	100.0	10	100.0	10	100.0	10	100.0
Shakespeare	42	40	95.2	41	97.6	41	97.6	41	97.6
Shaw	10	9	90.0	8	80.0	8	80.0	8	80.0
Stevenson	10	7	70.0	6	60.0	5	50.0	6	60.0
Stockton	10	9	90.0	8	80.0	8	80.0	8	80.0
Tolstoy	15	8	53.3	7	46.7	6	40.0	7	46.7
Twain	14	13	92.9	12	85.7	12	85.7	13	92.9
Verne	10	10	100.0	10	100.0	10	100.0	10	100.0
Wake	9	6	66.7	9	100.0	9	100.0	9	100.0
Warner	10	8	80.0	9	90.0	9	90.0	9	90.0
Wells	10	9	90.0	8	80.0	8	80.0	8	80.0
Wilde	7	2	28.6	3	42.9	2	28.6	2	28.6
Wodehouse	23	22	95.7	20	87.0	21	91.3	21	91.3
Yonge	10	8	80.0	7	70.0	8	80.0	9	90.0
Total	634	543	85.6	527	83.1	528	83.3	534	84.2

Table 4: Example ranked lists (top 5) for five works of Shakespeare; markers are function words only.

Rank	Sh139	Sh149	Sh155	Sh163	Sh166
1	Sh166	Sh165	Sh128	Sh162	Sh139
2	Sh145	Sh21	Sh162	Sh166	Sh148
3	Sh148	Sh29	Sh167	Sh169	Sh147
4	Sh147	Sh164	Sh147	Sh23	Sh145
5	Sh155	Sh22	Sh164	Sh168	Sh155

Master and Man

The hypothesis that the works of Shakespeare were written by someone else has been argued for hundreds of years.[5] As a preliminary investigation into whether our methods could throw any light on the debate, we included in Gutenberg634 the works of major playwrights of Shakespeare's time: William Shakespeare, Francis Beaumont & John Fletcher (whose works are co-authored), Ben Jonson, and Christopher Marlowe. By examining the extent to which these works were consistent with each other, and whose works matched to whose, we speculated that we might discover some evidence pointing in one direction or the other.

An admitted weakness of such an investigation is that our tools are not particularly sophisticated. These texts have been subjected to intensive literary analysis for several centuries and it would be foolhardy of us to suppose that a straightforward statistical analysis would lead to dramatic revelations. Nonetheless, it is our belief that patterns of writing are not easily mimicked or disguised. Should an author be pretending to be Shakespeare, for example, we would hope to observe inconsistencies in the statistical character of Shakespeare's works.

Another caveat is that we have not spread our net wide. Many candidates have been proposed as the authors of the works of Shakespeare; however, of these, only the works of Marlowe were available to us in a suitable form.

To examine this question of authorship, consider first the search experiments reported earlier, in which function words were used as markers. We now examine the ranked lists for selected books by each of these four cases, Shakespeare, Beaumont & Fletcher, Marlowe, and Jonson. For simplicity, we use a shorthand to indicate the writers in the following tables: "Sh", "BF", "Ma", and "Jn". Each notation is followed by a number, derived from filenames in Gutenberg634, so that for example Sh165 represents *Timon of Athens*.

In Table 4, we have listed the authorship of the top 5 retrieved books for each of five of Shakespeare's texts. For Shakespeare, we found an extremely high consistency of writing, with, overall, 203 of 210 top-5 listings being correct.

In Tables 5, 6, and 7 we show the ranked lists for Beaumont & Fletcher Marlowe, and Jonson. These results are rather less consistent than for Shakespeare, perhaps unsurprisingly given that there are far fewer training texts for these authors.

The best case is that of Beaumont & Fletcher, where only six of the 25 documents are mismatches. The cases of Marlowe and Jonson are more intriguing. Marlowe's rankings are dominated by the works of Shakespeare, with 17 of the 25 matches. Jonson is hardly better, with Shakespeare giving 14 of the 25 matches. In both cases the actual works of the author are not prominent. Note too that all but one of the 100 matches is by a playwright of this era—they may be conflated with each other, but there is no doubting when these plays were written.

[5]Some say that the proposition was first put by Edward Blount in 1623, others cite Queen Elizabeth I. See for example any number of web pages that are returned for the query "marlowe shakespeare". We hesitate to endorse them but they are certainly entertaining. Results for "shakespeare authorship" are also of interest.

Table 5: Example ranked lists for works of Beaumont & Fletcher; markers are function words only.

Rank	BF19	BF20	BF21	BF22	BF23
1	BF24	BF21	BF20	BF21	BF20
2	BF23	BF23	BF22	BF20	BF24
3	Sh149	BF22	BF23	BF23	BF21
4	Sh165	BF24	BF24	BF24	BF22
5	Sh159	Jn7	Jn8	BF19	Jn8

Table 6: Example ranked lists (top 5) for works of Marlowe; markers are function words only.

Rank	Ma11	Ma12	Ma13	Ma14	Ma17
1	Sh166	Ma13	Ma14	Ma13	Sh166
2	Sh163	Sh139	Sh139	Sh139	Sh139
3	Sh148	Ma14	Ma12	Ma12	Sh147
4	Sh139	Ma17	Sh166	Sh166	Sh155
5	Sh169	other	Sh147	Sh147	Sh148

So, does the evidence suggest that Marlowe wrote Shakespeare?

The circumstances of Marlowe's death, whether in a tavern or in an assassination, have been debated for longer than the question of authorship of the works of Shakespeare. Some people argue that Marlowe faked his death and used "Shakespeare" as his pen name to continue writing afterwards. However, our results do not suggest a particular relationship between the works of Marlowe and Shakespeare.

It is true that plays by Marlowe tend to retrieve plays by Shakespeare, as seen in Table 6. However, the evidence becomes weaker when we compare Table 4 and Table 6 in detail. Sh139 appears five times in five searches in Table 6. Given the hypothesis that the true author for this book is Marlowe, it should occasionally retrieve books by Marlowe. However, as can be seen in Table 4, when Sh139 is used as a query, no works of Marlowe are retrieved. Sh166 and Sh147 share the same properties—none of these retrieve books by Marlowe. The fact that Jonson's works also match those of Shakespeare further suggests that the similarity with Marlowe may be a matter of period rather than authorship.

The positive leave-one-out experiments are also indicative. In these experiments, the plays of Marlowe and Jonson are never misattributed. To some extent this may be due to experimental design—the presence of Shakespeare's plays in the negative examples is watered down by the great volume of nineteenth-century text. However, in the negative leave-one-out experiments, the works of both Marlowe and Jonson are usually attributed to Shakespeare, while those of Beaumont & Fletcher are occasionally attributed to Shakespeare. That is, the rate of false matches is extremely high, and the works of these authors cannot be distinguished. At the same time, there is no particular evidence that the works of any of these authors has unusually high similarity to that attributed to Shakespeare. Taking these considerations together we see no evidence in our experiments to support the hypothesis that Marlowe wrote Shakespeare.

Table 7: Example ranked lists (top 5) for works of Jonson; markers are function words only.

Rank	Jn1	Jn5	Jn7	Jn8	Jn9
1	Jn8	Jn7	Jn5	Sh142	Jn8
2	Sh162	Sh168	Jn2	Sh167	Jn2
3	Sh28	Jn1	Sh168	Jn2	Sh147
4	Sh142	Sh8	Jn8	Jn7	Sh139
5	Sh155	Sh156	Sh167	Jn1	Sh26

5 All's Well That Ends Well

We have explored the effectiveness of authorship attribution on works of literature. Using a collection of 634 works derived from the Gutenberg project, our experiments have shown that positive leave-one-out classification can be highly effective, with accuracy of over 85%. Negative leave-one-out experiments, although admittedly incomplete, were even more accurate. Search-based attribution was less successful, but still achieved accuracy of 74%. Not only, then, do these results confirm that authors do indeed have an idenfiable writing style, but they confirm that simple markers suffice to identify a particular author.

The best results used function words as markers of style; part-of-speech tags were reasonably effective, but were, we believe, undermined by the fact that tagging is an error-prone process. Tagging tends to fail on text with unusual constructions, and such constructions tend to be indicative of style. In contrast, extraction of function words is straightforward.

Most of our experiments used whole documents as queries. Use of fragments of documents was less successful, with 1000 words being clearly insufficient. Fragments of 10,000 words—somewhat over a tenth of a typical book—allowed correct attribution in over 50% of cases. This result is consistent with our previous exploration of AA on news articles, which are much shorter than books; the accuracy of classification is much better than random, but is insufficient on its own to definitively determine authorship.

These experiments allowed us to examine why attribution sometimes fails. The pattern of errors suggests that a key cause is a lack of distinct style in some texts, such as translated books. That is, some of the failures are due to properties of the works rather than shortcomings of the attribution method. The experiments also allowed us, in a small way, to revisit the question of the authorship of Shakespeare. We did not discover evidence that these works were written by Marlowe.

A limitation of our experiments was that the sources were somewhat mixed, and time prohibited creation of the larger pool of texts that we would have like to have used—each text required some manual editing to remove non-author material. The bulk of the text was from the nineteenth century, but a fraction was much older. Nonetheless, results were highly successful, and provide strong confirmation of the ability of simple statistical methods to accurately identify authorship.

Acknowledgements.

This work was supported by the Australian Research Council.

References

Baayen, H., Halteren, H. V., Neijt, A. & Tweedie, F. (2002), 'An experiment in authorship attribution',

Baayen, H., Halteren, H. V. & Tweedie, F. (1996), 'Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution', *Literary and Linguistic Computing* **11**(3), 121–132.

Bekkerman, R., El-Yaniv, R., Tishby, N. & Winter, Y. (2003), 'Distributional word clusters vs. words for text categorization', *J. Mach. Learn. Res.* **3**, 1183–1208.

Benedetto, D., Caglioti, E. & Loreto, V. (2002), 'Language trees and zipping', *The American Physical Society* **88**(4).

Binongo, J. N. G. (2003), 'Who wrote the 15th book of Oz? an application of multivariate statistics to authorship attribution', *Computational Linguistics* **16**(2), 9–17.

Burrows, J. (1987), 'Word patterns and story shapes: the statistical analysis of narrative style', *Literary and Linguistic Computing* **2**, 61–70.

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

Chen, S. F. & Goodman, J. (1996), An empirical study of smoothing techniques for language modeling, *in* A. Joshi & M. Palmer, eds, 'Proc. 34th Annual Meeting of the Association for Computational Linguistics', Morgan Kaufmanns, pp. 310–318.

Diederich, J., Kindermann, J., Leopold, E. & Paass, G. (2003), 'Authorship attribution with support vector machines', *Applied Intelligence* **19**(1-2), 109–123.

Fung, G. (2003), The disputed federalist papers: Svm feature selection via concave minimization, *in* 'Proc. 2003 Conf. on Diversity in Computing', ACM Press, pp. 42–46.

Gabrilovich, E. & Markovitch, S. (2004), Text categorization with many redundant features: Using aggressive feature selection to make svms competitive with c4.5, *in* 'Proc. 21st Int. Conf. on Machine learning', ACM Press.

Goodman, J. (2002), 'Extended comment on language trees and zipping'.

Hiemstra, D. (2002), Term-specific smoothing for the language modeling approach to information retrieval: the importance of a query term, *in* 'Proc. 25th ACM SIGIR Conf. on Research and Development in Information Retrieval', ACM Press, pp. 35–41.

Holmes, D. I., Robertson, M. & Paez, R. (2001), 'Stephen Crane and the New York Tribune: A case study in traditional and non-traditional authorship attribution', *Computers and the Humanities* **35**(3), 315–331.

Juola, P. & Baayen, H. (2003), 'A controlled-corpus experiment in authorship identification by cross-entropy', *Literary and Linguistic Computing* .

Kaster, A., Siersdorfer, S. & Weikum, G. (2005), Combining text and linguistic doument representations for authorship attribution, *in* 'SIGIR workshop: Stylistic Analysis of Text For Information Access'.

based measure for verification of text collections and for text categorization, *in* 'Proc. 26th ACM SIGIR Conf. on Research and Development in Information Retrieval', ACM Press, pp. 104–110.

Khmelev, D. V. & Tweedie, F. (2002), 'Using markov chains for identification of writers', *Literary and Linguistic Computing* **16**(4), 229–307.

Koppel, M. & Schler, J. (2004), Authorship verification as a one-class classification problem, *in* 'Proc. 21st Int. Conf. on Machine Learning', ACM Press.

Kukushkina, O., Polikarpov, A. & Khmelev, D. (2000), 'Using literal and grammatical statistics for authorship attribution'.

Lafferty, J. & Zhai, C. X. (2001), Document language models, query models, and risk minimization for information retrieval, *in* 'Proc. 24th ACM SIGIR Conf. on Research and Development in Information Retrieval', ACM Press, pp. 111–119.

Li, J. X., Zheng, R. & Chen, H. C. (2006), 'From fingerprint to writeprint', *ACM Commun.* **49**(4), 76–82.

Li, T., Zhu, S. H. & Ogihara, M. (2003), Efficient multi-way text categorization via generalized discriminant analysis, *in* 'Proc. 12th Int. Conf. on Information and Knowledge Management', ACM Press, pp. 317–324.

Masuyama, T. & Nakagawa, H. (2004), Two step pos selection for svm based text categorization, *in* 'IEICE Transaction on Information System', Vol. E87-D.

Pol, M. S. (2005), A stylometry-based method to measure intra and inter-authorial faithfulness for forensic applications, *in* 'SIGIR workshop: Stylistic Analysis of Text For Information Access'.

Quinlan, R. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann.

Sarkar, A., Roeck, A. D. & Garthwaite, P. H. (2005), Term re-occurrence measures for analyzing style, *in* 'SIGIR workshop: Stylistic Analysis of Text For Information Access'.

Scholkopf, B. & Smola, A. J. (2002), *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*, MIT Press.

Stamatatos, E., Fakotakis, N. & Kokkinakis, G. (1999), Automatic authorship attribution, *in* 'Proc. 9th Conf. of the European Chapter of the Association for Computational Linguistics', pp. 158–164.

Stamatatos, E., Fakotakis, N. & Kokkinakis, G. (2001), 'Computer-based authorship attribution without lexical measures', *Computers and the Humanities* **35**(2), 193–214.

Witten, I. H. & Frank, E. (2000), *Data Mining: Practical Machine Learning Tools with Java Implementations*, Morgan Kaufmann.

Zhai, C. X. & Lafferty, J. (2001), A study of smoothing methods for language models applied to ad hoc information retrieval, *in* 'Proc. 24th ACM SIGIR Conf. on Research and Development in Information Retrieval', ACM Press, pp. 334–342.

Zhai, C. X. & Lafferty, J. (2002), Two-stage language models for information retrieval, *in* 'Proc. 25th ACM SIGIR Conf. on Research and Development in Information Retrieval', ACM Press, pp. 49–56.

ing methods for language models applied to information retrieval', *ACM Transaction on Information System* **22**(2), 179–214.

Zhao, Y. & Zobel, J. (2005), Effective authorship attribution using function word, *in* 'Proc. 2nd AIRS Asian Information Retrieval Symposium', Springer, pp. 174–190.

Zhao, Y. & Zobel, J. (n.d.), 'Authorship search in large document collections'. Manuscript in preparation.

Zhao, Y., Zobel, J. & Vines, P. (2006), Using relative entropy for authorship attribution, *in* 'Proc. 3rd AIRS Asian Information Retrieval Symposium', Springer. to appear.

Zobel, J. & Moffat, A. (2006), 'Inverted files for text search engines', *ACM Computing Surveys* . To appear.