Authorship Attribution Using Lexical Attraction

by

Corey M. Gerritsen

Submitted to the Department of Electrical Engineering and Computer Science

in Partial Fulfillment of the Requirements for the Degrees of

Bachelor of Science in Computer Science and Engineering

and Master of Engineering in Electrical Engineering and Computer Science

at the Massachusetts Institute of Technology

May 21st, 2003

Author _____

Department of Electrical Engineering and Computer Science
May 21st, 2003

Certified By _____

Patrick Henry Winston
Ford Professor of Artificial Intelligence and Computer Science
Thesis Supervisor

Accepted By _____

Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Authorship Attribution Using Lexical Attraction

by Corey M. Gerritsen

Submitted to the
Department of Electrical Engineering and Computer Science

May 21st, 2003

In Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Computer Science and Engineering
and Master of Engineering in Electrical Engineering and Computer Science

ABSTRACT

Authorship attribution determines who wrote a text when it is unclear who wrote the text. Some examples are when two or more people claim to have written something or when no one is willing (or able) to say that he or she wrote the piece. In order to further the tools available for authorship attribution, I introduced lexical attraction as a way to distinguish authors. I implemented a program called StyleChooser that determines the author of a text, based on Yuret's lexical attraction parser. StyleChooser, once trained on a set of authors, determines how much information is redundant under each author model. Dividing by the number of words in the test text and by the log of the number of words used to train the model gives a metric used to rank the known authors in order of likelihood that they wrote the text in question. I then tested StyleChooser and analyzed the results. When tested with knowledge of 62 authors on 369 texts by those authors, my program had an accuracy of 75%, while the right author ranked in the top three authors 86% of the time. The closeness of a few authors shows that StyleChooser does a better job of differentiating between styles in a broader sense than between authors. A program that differentiates between styles could be used for style differentiation, style based searching, and even better human/computer interaction.

Thesis Supervisor: Patrick Winston
Title: Ford Professor of Artificial Intelligence and Computer Science

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The author wishes to thank Deniz Yuret, for inspiring him; Patrick Winston, for the ideas that kept him interested; his parents, for giving him the tools to get where he is today; and Sara Elice, just for existing.

1. Introduction

Authorship attribution empowers governments and institutions to give credit where credit is due, be it for scholarly works or for terrorist manifestos. Literary scholars have been and continue to be the foremost experts on authorship, but stylometry, the statistical analysis of style in literature, has expanded as a field in the past century. Stylometry attempts to capture an author's style using quantitative measurements of various features in the text such as word length or vocabulary distributions. Many stylometric studies have measured word dependencies as a feature of an author's style using language models that restrict what words a given word can depend upon.

A lexical attraction based language model is a probabilistic language model in which each word can depend upon any other word within the same sentence. I believe that an author chooses a word not based solely upon the words that come before it, as *n-gram* models presume, but instead based on all the other words in the sentence. Therefore, lexical attraction should lead to better quantitative authorship attribution than previous word dependency models which further restricted the dependencies between words.

I built a program, StyleChooser, that processes texts to collect lexical attraction data for given authors, and then uses the statistical data in an attempt to determine the author of a given text. StyleChooser uses lexical attraction as a metric to distinguish between the candidate authors, giving a ranking from which a degree of confidence can be established. StyleChooser has trouble distinguishing between certain authors, which I attribute to the authors having very similar styles. My program could be used to distinguish between

styles very well. StyleChooser does not need special processing to be performed on the input texts, and can therefore be used for any machine readable text, regardless of language.

The rest of Section 1 explains the motivation behind using lexical attraction for authorship attribution. Section 2 provides background information about stylometry and lexical attraction. Section 3 describes my program, StyleChooser, and the tests I performed with it. Section 4 gives the results of the experiments described in Section 3. Section 5 is a discussion of the results, with suggestions for future work, and Section 6 summarizes the contributions of this thesis.

1.1 The Importance of Authorship Attribution

Authorship attribution, as the name implies, involves determining the author of a disputed work. Authorship attribution is used in the purely scholarly sense: to decide whether the works of Shakespeare were written by one man, or to determine who wrote the twelve *Federalist Papers* that were written anonymously. Authorship attribution also comes up in criminal investigation: Ted Kaczynski was targeted as a primary suspect in the Unabomber case because authorship attribution methods determined that he could have written the Unabomber's manifesto (Foster, 2000).

Authorship attribution can also be used to stop plagiarism, which has become more of a problem with the internet boom. Anyone can take a copy of someone else's text and put it on the web with his or her own name on it. Authorship attribution methods are important to determine who deserves recognition for the work.

Understanding authorship and how to tell the difference between authors could lead to a better understanding of language. Because language is a key feature of human intelligence, by understanding the regularities in various authors' writing and styles we may come closer to understanding how the mind works.

1.2 Why Lexical Attraction?

Lexical attraction is the likelihood of two words in a sentence being related. One of the foremost parts of style is word choice. Previous stylometry models that use word choice have looked at vocabulary distributions, function word usage, and *n-gram* word dependencies, but not at the relation between words throughout a sentence (Holmes, 1994, Diedrich et al, 2000).

Lexical attraction gives a dependency linking for each sentence that captures how words are chosen based on each other throughout a sentence. Because word choice is such a large part of style, lexical attraction can be used to differentiate between different styles. Assuming that each other has his or her own style, lexical attraction can therefore be used for authorship attribution. I hypothesize that lexical attraction can be used to differentiate between different styles, and through this differentiation do a decent job of authorship attribution.

2. Background

Before discussing StyleChooser and what I did with it, I give some background behind the field of stylometry and lexical attraction models. Section 2.1 explains stylometry in more detail, and section 2.2 describes lexical attraction language models and how they work.

2.1 Stylometry

Because stylometry is the statistical analysis of literary style, the main assumption behind stylometry is that authors make certain subconscious and conscious choices in their writing. The hypothesis is that an author will not be able to control the subconscious choices, and therefore the same choices will be made across an author's work. If those choices can be recognized, they can be used to identify the author's work. Stylometrists have tried to identify various features of an author's writing that remain constant throughout that author's work.

Some of the features that have been used in stylometry include average sentence length, average syllables per word, average word length, distribution of parts of speech, function word usage, the Type-Token ratio, Simpson's Index, Yule's Characteristic K, entropy, word frequencies, and vocabulary distributions (Holmes, 1994). Some models that have been used in stylometry include n-grams, feature counts, inductive rule learning, Bayesian networks, radial basis function networks, decision trees, nearest neighbor

classification, and support vector machines (Diedrich et al., 2000). There have also been studies using content analysis and neural networks (Holmes, 1997).

The straight probabilistic models that have been used, n-grams and feature counts, make little use of the interplay between words. The n-gram model comes close, making each word depend on the n previous words. But this doesn't pay attention to the fact that each word chosen by the author affects the other words chosen by the author, throughout a sentence, paragraph, or even an entire novel. Straight word counts catch some of the aspects of word choice, but not the interplay between words.

2.2 Lexical Attraction Model of Language

The lexical attraction model, on the other hand, is a model that focuses on the interactions between words anywhere in a sentence. Yuret defines *lexical attraction models* as the class of language models that are based on the assumption that each word depends on one other word in the sentence, but not necessarily an adjacent word (Yuret, 1998). Yuret connects this dependence to a syntactic relation, but for authorship attribution, this connection is unnecessary. Lexical attraction just means that each word in the sentence is linked to one other word in the sentence in a dependency structure.

The lexical attraction model I use in my program can be directly attributed to Deniz Yuret. A full explanation of this model and proofs of the following statements are provided in Yuret's thesis (Yuret, 1998):

- each sentence has a link structure that indicates which words are linked,

- the total information in a sentence = the information in the words
  + the information in the dependency structure
  – the mutual information captured in syntactic relations,
- the average information of a word is independent of the language model,
- information in the dependency structure is linear in the number of words in the sentence,
- and therefore to compare the entropy in a sentence under lexical attraction based language models, only the mutual information in the word dependencies needs to be compared.

The important parts are that how much information is in the sentence depends upon the language model that is in effect, and that the efficiency of different language models can be compared by paying attention only to the mutual information in the relations between words.

3. Methods

I implemented a program, StyleChooser that determines who wrote a piece of text from candidate authors with available training data. To determine who wrote the text, first StyleChooser trains a language model for each candidate author, using the training data for that author. Next, StyleChooser calculates the amount of information that exists in the word dependencies using each candidate's author model. The language model that causes the most mutual information in the word links in the text is declared the most likely model, and the corresponding author is determined to have written the text.

The training sequence is described in 3.1, Creating Author Models.

Section 3.2, A Metric for Determining Authorship, describes the testing sequence and explains why mutual information is a good metric for authorship attribution.

Section 3.3, Tests for Success, describes tests to determine if my program accurately differentiates between authors and tests to determine how factors such as training set size and test set size affect the accuracy of my program.

3.1 Creating Author Models

In order to determine what author wrote a text, my program needs to first create a language model for each candidate author. The training algorithm for my program is Yuret's bootstrapping acquisition method (Yuret, 1998). This learning mechanism interweaves learning and processing so that information collected so far can affect the next iteration of learning, to speed up the training process.

The goal of the training algorithm is to find the dependency structure for each sentence that maximizes the mutual information in the links in the sentence (Yuret describes the process as trying to assign the sentence the highest probability, Yuret, 1998). An optimal algorithm, that Yuret describes, runs in $O(n^5)$ time, and Yuret suggests and uses an approximation algorithm, called the Link-Sentence algorithm, that runs in $O(n^2)$. I use Yuret's Link-Sentence algorithm, the pseudo code of which is in Figure 1. While the approximation algorithm might not find the absolute best dependency structure for a given sentence, every action it performs moves to a better dependency structure. Using the approximation algorithm instead of an optimal algorithm could have an affect on the accuracy of StyleChooser; I will discuss this further in Section 5.3, Inaccuracy.

```
Link-Sentence(S)
1  for j = 1 to Length(S)
2       for i = j - 1 downto 1
3            last = Pop(Right-Links(i),stack)
4            minlink[i] = Min(last,minlink[Right-Word(last)])
5            if MI(S[i],S[j]) > 0
6                 and MI(S[i],S[j]) > MI(minlink[i])
7                 and ∀s : MI(S[i],S[j]) > MI(stack[s])
8            then ∀s : Unlink(stack[s])
9                 Reset(stack)
10           Unlink(minlink[i])
11           minlink[i] = Link(S[i],S[j])
12           Push(Left-Links(i),stack)
```

Figure 1: Link-Sentence Pseudocode

In order to find dependency structures and to encode each sentence, for any word x and for any word y the model needs to keep track of the number of times that x is linked to y, the number of times that x is linked to any word, and the number of times that any word is linked to y (note that x can be the same word as y.) It also needs to keep track of the total number of links that have been made. With these counts, the mutual information (MI) between the words x and y can be computed, if *n(x,y)* is the number of times x is linked to y, N is the total number of links in the model so far, and * stands for any word (Yuret, 1998):

$$MI(x,y) = \log_2\left(\frac{n(x,y)N}{n(x,*)n(*,y)}\right).$$

The counts are updated every time the training algorithm finishes with a sentence, so that the results of the n^{th} sentence can be used to improve the results of the $(n+1)^{th}$ sentence in the training set.

Yuret's thesis provides further explanation of how the training algorithm works and in-depth examples of the algorithm at work.

3.2 A Metric for Determining Authorship

Once author models for each candidate author have been trained, StyleChooser can figure out who wrote anonymous texts. To determine the author that best matches an anonymous text, StyleChooser takes each candidate author one at a time: for each candidate author, StyleChooser uses the same basic algorithm used in the training stage to determine the dependency structure in each sentence, except that it does not update any counts in the author language models. As it processes the anonymous text, StyleChooser keeps track of the amount of mutual information resident in the word links using the current author model.

The author that wrote the anonymous text has a language model that needs the least amount of information to represent the anonymous text. As discussed earlier, comparing the information needed to represent a text as the language model is changed can be simplified to comparing only the mutual information in links in the text – the information in the dependency structure can be ignored because it is a constant size and the information needed to encode the words can be ignored because it is not a part of the lexical attraction language model and is therefore the same across different models. Therefore, the total

mutual information in all the links in the anonymous text under each model is used as a metric to determine which author is the best match and how good of a match that author is.

If the total mutual information is divided by the number of words in the anonymous text, the resultant number can be compared between different runs of StyleChooser to give relative certainties between different authorship attribution runs (i.e., runs on different anonymous texts).

Another way to think of this metric is to assume we are trying to compress the unknown text using the lexical attraction models of each author, as described by Bach and Witten (Bach and Witten, 1999). Bach and Witten use lexical attraction to compress large corpora by training over the text and then storing the text using the minimum amount of information necessary by eliminating the shared information in dependency links with a minimum length encoding. In other words, the mutual information returned by using an author model on a text is the amount of information that would be saved if the author model were used to compress the text. The author that compresses the text the most is the author that wrote the text.

3.3 Tests for Success

To determine how well lexical attraction works as an authorship attribution model, it is important to see how this program fares against a large scale accuracy test, how the training set size affects the outcome of determination, and how the test set size affects the outcome of determination.

The most common first test for an authorship attributor is to determine who wrote eleven of the *Federalist Papers* that were written by either James Madison or Alexander Hamilton, but for which there has been a long standing dispute as to who wrote them. Other authorship attributors have found that Madison wrote all twelve of the disputed texts (Mosteller and Wallace, 1964; Bosch and Smith, 1998). StyleChooser trained an author model for both Hamilton and Madison based upon the *Federalist Papers* that they wrote, for Hamilton: 1, 6-9, 11-13, 15-17, 21-36, 59-61, and 65-85; and for Madison: 10, 14, 37-48, and 58. Each of the disputed papers, 49-57 and 62-63, were then tested to see which author was a better fit according to StyleChooser.

In order to more completely test StyleChooser, 431 different texts by 62 different authors (at least two texts per author) were downloaded from the Project Gutenberg[1] website. The meta data (distribution information, etc.) for each text was removed, leaving just the text of each text. StyleChooser was trained with one text by each author, chosen randomly, and then the remaining 369 texts were given to StyleChooser one at a time to determine which of the 62 authors wrote each text. The texts used for training, henceforth referred to as Set A, are listed in Table 6 in Appendix A, and the texts used for testing, henceforth referred to as Set B, are listed in Table 7 in Appendix A.

To determine how training set length affects accuracy, another test was performed in which the test texts from Set B were used to train author models, and then StyleChooser determined the authors of the texts in Set A.

[1] Project Gutenberg can be found at http://promo.net/pg/

Another test to ascertain the effect of training set length on the value of the mutual information per word was performed, in which author models were created from *Great Expectations*, by Charles Dickens, and *Emma*, by Jane Austen, for 500 word increments of those books. StyleChooser then calculated the mutual information per word for each of these author models on *Mansfield Park*, also by Jane Austen. Due to the results of this test, the metric for comparing two authors was changed from the mutual information per word to the mutual information per word divided by the logarithm of the length of the training text, and previous accuracy tests were performed again with the new metric.

Another aspect that could affect the value of the mutual information per word is the length of the test text of unknown authorship. After training author models from the entirety of *Great Expectations* and *Emma*, *Mansfield Park* was split into 500 word portions and StyleChooser calculated the mutual information per word for Dickens and Austen on test sets of multiples of 500 words.

Based on the results of this last test, the accuracy test involving training with Project Gutenberg Set A and testing with Set B was run again, except only the first 15,000 words of each text in Set B were used in the determination process.

4. Results

Before running any other tests, I first tried my authorship attributor on the eleven anonymous *Federalist Papers*, which could have been written by either Alexander Hamilton or James Madison. My program agreed with previous works that agree that Madison wrote all eleven of the disputed papers (Mosteller and Wallace, 1964, Bosch and Smith, 1998). The numerical results are shown in Table 1. Having established that StyleChooser agrees with other methods, I moved on to further tests.

Table 1: Mutual Information per word for Hamilton and Madison for each of the disputed *Federalist Papers*

Paper #	Author	MI/word
49		
	Hamilton	2.63306
	Madison	2.564283
50		
	Hamilton	2.47404
	Madison	2.441458
51		
	Hamilton	2.660655
	Madison	2.532428
52		
	Hamilton	2.753937
	Madison	2.518712
53		
	Hamilton	2.712992
	Madison	2.470682
54		
	Hamilton	2.542816
	Madison	2.408824
55		
	Hamilton	2.686316
	Madison	2.538884
56		
	Hamilton	2.751647
	Madison	2.563197
57		
	Hamilton	2.620755
	Madison	2.473906
62		
	Hamilton	2.628089
	Madison	2.46058
63		
	Hamilton	2.636893
	Madison	2.543783

Next, I ran the large scale accuracy test described earlier (62 training texts listed in Appendix A in Table 6, 369 test texts listed in Appendix A in Table 7) to see how accurate my authorship determiner was. For each test text, I collected the rank at which StyleChooser placed the correct author. A summary appears in Table 2, while the specific ranking of the correct author for each text appears in Table 8 in Appendix B.

Table 2: number of texts that had their correct author ranked at each rank when training with Set A (one text per author) and testing with Set B (the rest of the texts/variable number per author)

Ranking	# texts	Ranking	# texts	Ranking	# texts	Ranking	# texts
1	161	17	1	33	0	48	0
2	24	18	2	34	3	49	0
3	18	19	4	35	1	50	0
4	22	20	2	36	0	51	0
5	18	21	3	37	1	52	0
6	20	22	2	38	2	53	0
7	12	23	2	39	0	54	0
8	10	24	1	40	1	55	0
9	6	25	2	41	1	56	0
10	7	26	2	42	0	57	0
11	6	27	0	43	0	58	0
12	6	28	2	44	0	59	0
13	2	29	1	45	1	60	0
14	11	30	2	46	0	61	0
15	4	31	1	47	0	62	0
16	3	32	2				

The results of the incremental text size testing with *Great Expectations*, *Emma*, and *Mansfield Park* are shown in Figure 2 and Figure 3. Figure 2 shows the effect that varying the training set size has on the amount of mutual information under a few author models. Figure 3 shows the effect that varying the test set size has on the amount of mutual information in the same situation. Figure 4 and Figure 5 show the mutual information as both the test text and training text sizes are varied. Cross sections of these graphs at the largest test text size give the data for Figure 2, and cross sections at the largest training set size make up the data in Figure 3.

Figure 2 shows two trend lines for the data, one for training with *Great Expectations* and one for training with *Emma*, that show a logarithmic relationship. Based on this empirical relationship, I changed the metric from mutual information per word to mutual information per word over log(training length), and repeated the first accuracy test. Results of training on Set A and testing on Set B with the new metric are listed in Table 3.

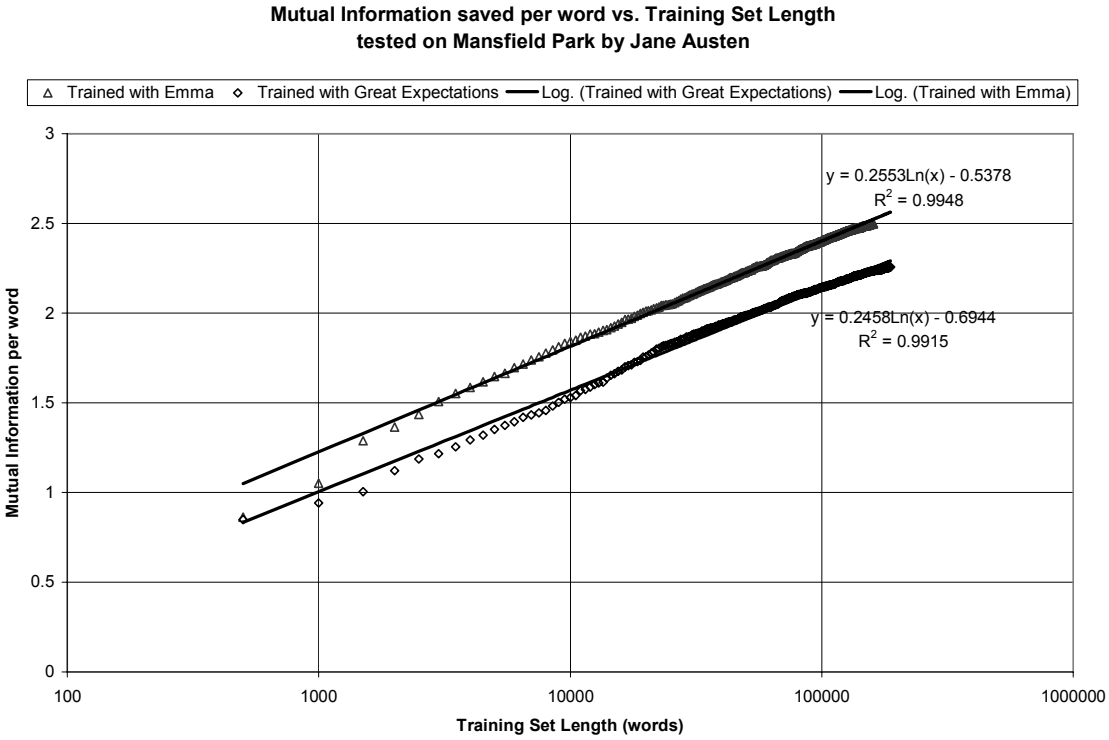

Figure 2: Effect of varying the training set size

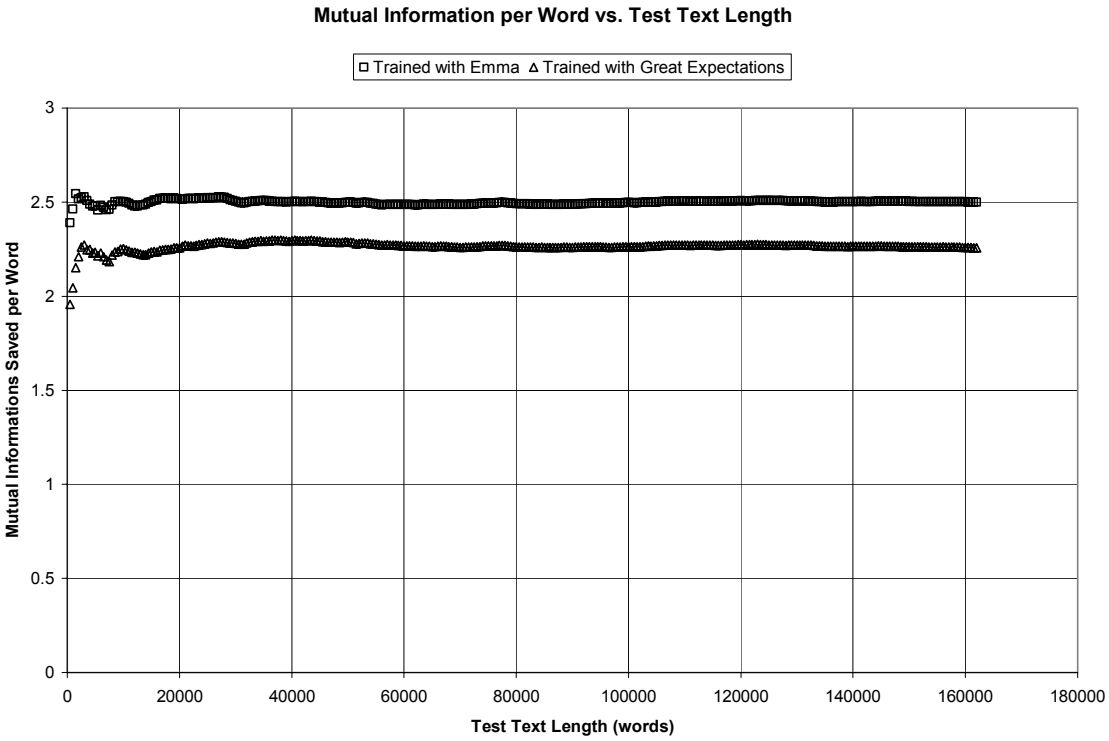

Figure 3: Effect of varying the test set size

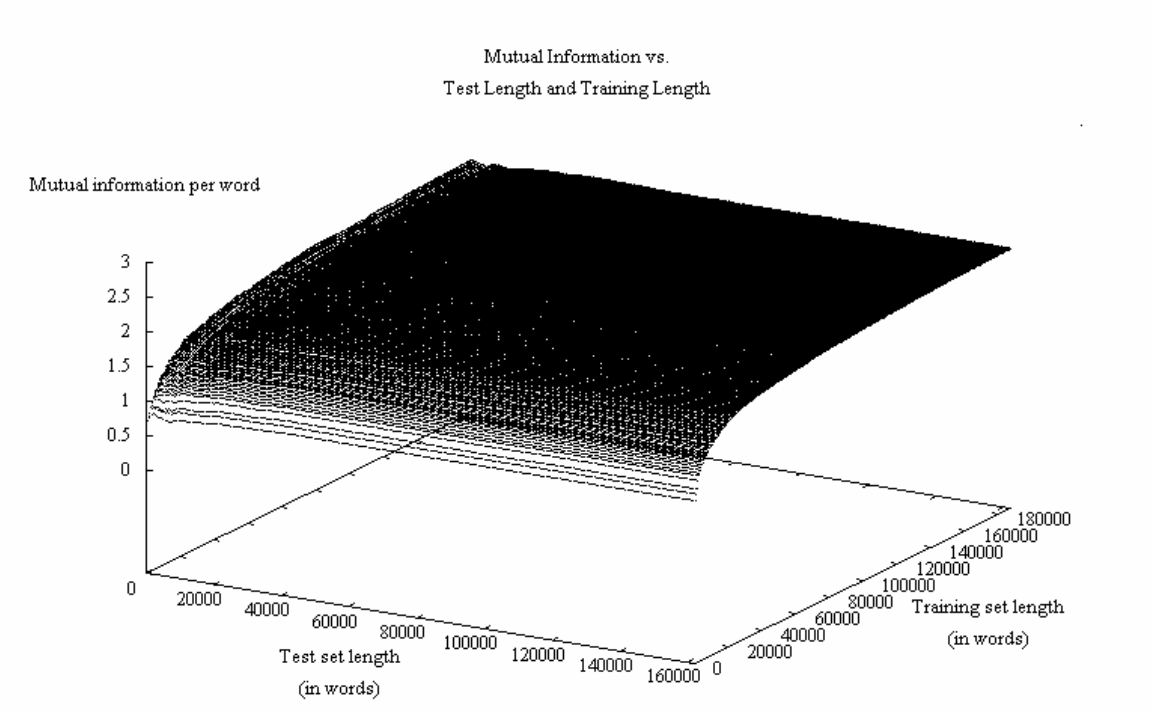

Figure 4: Full graph of the effects of changing the training and testing lengths.
Trained on *Great Expectations* and tested on *Mansfield Park*.

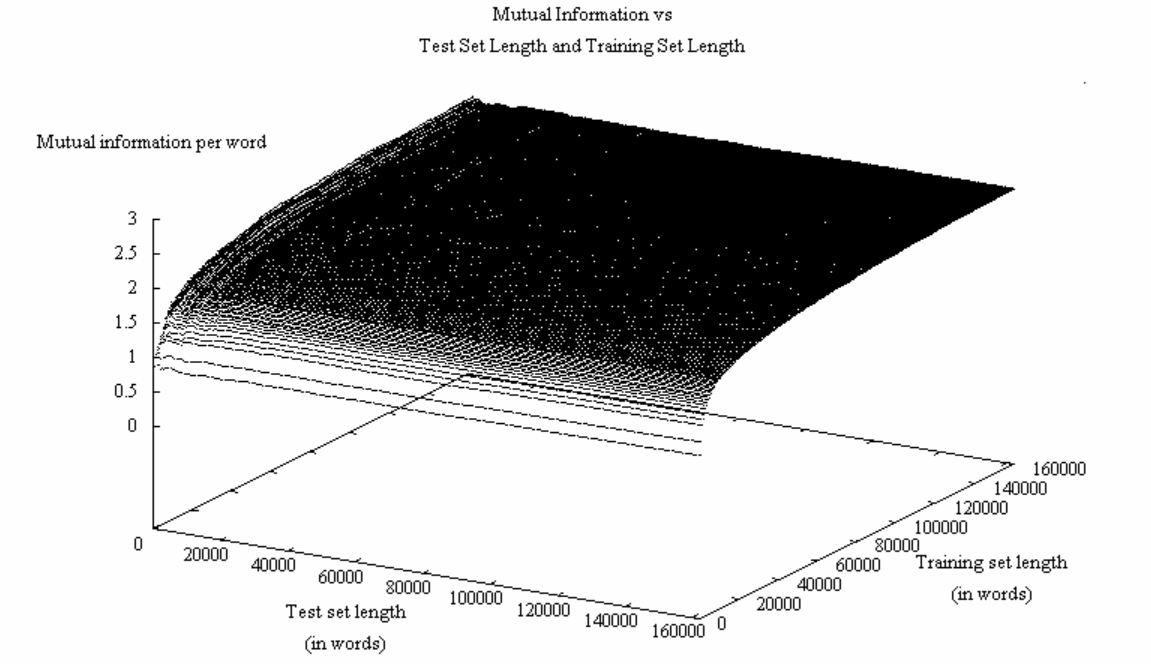

Figure 5: Full graph of the effects of changing the training and testing lengths.
Trained on *Emma* and tested on *Mansfield Park*.

In the graphs in Figure 3, it appears that at about 10,000 words of testing the average mutual information per word has leveled out and does not vary much (I will discuss this further in 5.1 Text Lengths). Based on this number, I decided to run another large scale accuracy test with Sets A and B from the project Gutenberg texts, only this time I only tested on the first 15,000 words of each test text. A summary of where StyleChooser ranked the correct authors is given in Table 4. Table 5 lists the results of the same experiment with the new metric, $\frac{MI}{length(test) \times \log(length(training))}$.

Table 3: Number of texts that had their correct author ranked at each rank when training with Set A and testing on Set B, with the metric of MI/(test set size)/log(training set size)

Ranking	# texts	Ranking	# texts	Ranking	# texts	Ranking	# texts
1	275	17	3	33	0	48	0
2	30	18	1	34	0	49	0
3	14	19	1	35	0	50	0
4	7	20	1	36	0	51	0
5	6	21	3	37	0	52	1
6	3	22	0	38	1	53	0
7	3	23	1	39	0	54	0
8	4	24	1	40	1	55	0
9	3	25	1	41	0	56	0
10	0	26	0	42	0	57	0
11	1	27	0	43	0	58	0
12	0	28	0	44	0	59	0
13	0	29	1	45	0	60	0
14	0	30	0	46	0	61	0
15	4	31	0	47	0	62	0
16	2	32	1				

Table 4: number of texts that had their correct author ranked at each rank when training with Set A and testing with the first 15000 words of each text in Set B

Ranking	# texts	Ranking	# texts	Ranking	# texts	Ranking	# texts
1	147	17	5	33	0	48	0
2	40	18	1	34	2	49	0
3	20	19	4	35	0	50	0
4	22	20	4	36	0	51	0

5	6	21	3	37	1	52	0
6	14	22	1	38	2	53	0
7	15	23	3	39	3	54	0
8	7	24	1	40	0	55	0
9	9	25	3	41	0	56	0
10	13	26	0	42	0	57	0
11	5	27	1	43	0	58	0
12	10	28	3	44	0	59	0
13	7	29	0	45	0	60	0
14	4	30	2	46	0	61	0
15	5	31	2	47	0	62	0
16	3	32	1				

Table 5: number of texts that had their correct author ranked at each rank when training with Set A, testing with the first 15000 words of each text in Set B, and using the MI/(words in test set)/log(words in training set) metric

Ranking	# texts	Ranking	# texts	Ranking	# texts	Ranking	# texts
1	267	17	0	33	0	48	0
2	32	18	2	34	0	49	0
3	14	19	1	35	0	50	0
4	9	20	0	36	1	51	0
5	5	21	2	37	1	52	0
6	5	22	1	38	0	53	0
7	3	23	0	39	0	54	0
8	2	24	1	40	1	55	0
9	3	25	0	41	0	56	0
10	5	26	0	42	1	57	0
11	1	27	2	43	0	58	0
12	3	28	0	44	0	59	0
13	1	29	1	45	1	60	0
14	2	30	0	46	0	61	0
15	0	31	1	47	0	62	0
16	0	32	1				

5. Discussion

In this section I will discuss the results from Section 4.

Section 5.1, Text Lengths, discusses the effect that changing the length of the test and training texts has on StyleChooser.

Section 5.2, Accuracy, conveys how well StyleChooser performed.

Section 5.3, Inaccuracy, talks about what kept StyleChooser from performing better.

Section 5.4, Future Work, explores future uses for StyleChooser.

5.1 Text Lengths

Figure 3 shows the effect that the length of the text of unknown authorship has on the mutual information per word. The graphs look like straight lines with a little random noise proportional to the inverse of the training set length. Each sentence in the training set is processed separately by StyleChooser, so it makes sense that different blocks of text in a text could differ in the amount of mutual information under a given author model. This is the entire reason why an average is being taken.

As the test size increases, the graphs level out. The fact that they level out implies that the average mutual information per word is fairly uniform over the entire test text. The question is how much input StyleChooser needs to come up with a stable answer. The value that each graph levels out to is the number that StyleChooser should use to compare how well different authors fit a disputed text. There is a closer look at the graphs of Figure

3 in Figure 6. For the tests shown in Figure 6, the mutual information per word has leveled out after only 15,000 words of test text. This implies that after 15,000 words by the same author, StyleChooser will return about the same average mutual information per word.

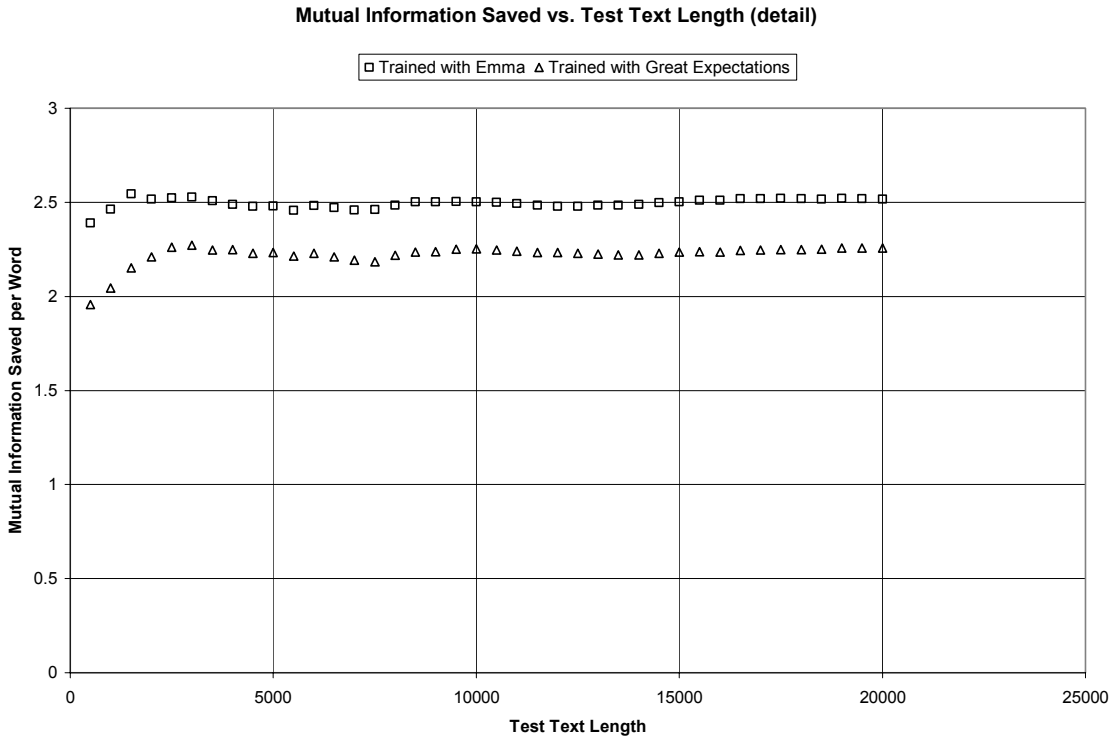

Figure 6: A closer look at the graph in Figure 3. The X-axis has been stretched out to get a better look at the variation in the MI close to x = 0.

Figure 2 shows the effect that training set length has on the mutual information per word. For both authors, the r^2 value is very high for a logarithmic best fit line. The stunning result is that the mutual information is proportional to the logarithm of the training set length. This suggests an immediate change to the metric in order to achieve independence from the training set size.

Why are training set size and mutual information logarithmically related? The equation for the mutual information is $MI(x,y) = \log_2(\frac{n(x,y)N}{n(x,*)n(*,y)})$. N in this equation

is the total number of links seen so far in training. This is the sum of the number of links in each sentence of the training set. Because the link dependency structure for a sentence is of size O(n), where n is the number of words in the sentence (Yuret, 1998), N is proportional to the size of the training set.

This leaves $\frac{n(x,y)}{n(x,*)n(*,y)}$, the number of times x is linked to y divided by the number of times x is linked to any word and the number of times any word is linked to y. For mutual information to be logarithmically related to the size of the training set, this quantity must not be independent of the size of the training set. When a word x appears in a sentence, it is likely to be linked to at most a constant number of words. So the number of times x is linked to another word is going to be proportional to the number of times it appears in the text. A similar argument can be made for the number of times any word is linked to y. For this quantity to be independent of the training text length, *n(x,y)* must be proportional to the product of the number of times x appears in the text and the number of times y appears in the text. This means that as words are used together, they continue to be used together at the same rate.

As the training set increases in size, the mutual information in a test text grows proportionally to the log of the size of the training set. This means that the length of the training set can be accounted for by StyleChooser to make comparisons between authors independent of training text length.

The end result of these tests is that lexical attraction works as an authorship attribution method regardless of the text sizes that can be obtained for training and testing. Because of the precise relationship between training size and mutual information,

differences in training size can be accounted for by StyleChooser. Except for at the very beginning of training, continued training of the same author does not have much effect on the average mutual information when divided by the log of the training length. Very little test data is needed also. It only takes about 15,000 words for the randomness inherent in any author's writing to stop having a noticeable effect on the outcome of testing. The sample size could be pushed lower, but if it is, the outcome of mutual information finding may be ruled by random eccentricities of the sentences being analyzed.

5.2 Accuracy

Without any controls on the lengths of texts for training or testing, StyleChooser correctly identified the authors of 161 out of 369 texts downloaded from Project Gutenberg, giving an accuracy of 43.6% in choosing the right author among 62 random authors (full results from the test are in Table 2). It may be a little unfair to force StyleChooser to differentiate between 62 different authors. If the determinations are limited to only two candidate authors, where one of the authors is the correct author, the accuracy becomes 91%. This can be determined from the ranking table by figuring out how many other authors each author was placed above for a particular text, adding these together, and dividing by the total number of pair-wise match-ups – the number of texts times the number of authors -1. Despite the great accuracy given using this method, it is slightly flawed: stylometric authorship attribution is meant to be used when it is difficult to figure out who the correct author of a text is, and there would never be a debate about whether it was Frederick Douglass or Charles Dickens who wrote a particular piece.

After seeing how the training set size affects the outcome of StyleChooser, I modified the metric to try to correct for differences in training length between author models. The results, listed in Table 3, show a greatly improved accuracy of 275 out of 369, or 74.5%. For a single feature of authorship, this is quite high, especially for allowing 62 candidate authors. When the test data was restricted to only 15,000 Words, the accuracy only lowered slightly, to 267 out of 369, or 72.4%.

5.3 Inaccuracy

Of course, 75% is not 100%, not even 90%. Why shouldn't lexical attraction allow for higher accuracy rates?

One answer could be that StyleChooser used a non-optimal algorithm for finding dependency structures in sentences. But not finding the best dependency structure just means that the model trades updating one count for updating another. Based on the fact that the mutual information is directly proportional to the log of the training set length, I believe that this slight degradation in the training would have an equal effect on all author models, and so should not affect StyleChooser's decision as to who wrote a text.

In my accuracy tests, StyleChooser has a few texts for which it ranks the correct author very low. There are a couple reasons this might happen. I chose the texts to be used for training and testing randomly. It is possible that the text used for training for a particular author did not match the style of his or her other work, or that a particular test text was unique to the author. Another possibility is that the style of an author changes

over time, so that an author model trained on a text written early in an author's career might be a bad match stylistically for a text written much later.

There are only four texts for which the correct author is ranked greater than 30: *Aaron's Rod* and *Women in Love*, by Lawrence, and *Little Lord Fontleroy* and *The Secret Garden*, by Burnett. StyleChooser performed worst of all on the two texts by Lawrence, implying that the text used to train Lawrence was not representative of his work. The texts by Burnett also don't fare very well. This could be because *A Lady of Quality*, the book used to train the model for Burnett, is a book meant for adults, while a lot of Burnett's works are children's novels. Lexical attraction depends on word choice, and audience has a large effect on word choice. The different audiences for Burnett's works probably changed her choices for words, which makes it hard to use lexical attraction to identify her as the author.

Another reason StyleChooser is not more accurate could be that word choice is just one element of style, and an element that can be very similar between authors. My test bed of 62 authors had a lot of similar authors in it. If I assume that for each author in my 62 author pool there are 2 other authors that can be considered stylistically close enough that being ranked below them is not necessarily a miss, then the accuracy rises to 86.4%. In this case, StyleChooser is no longer differentiating between authors, but instead between styles, where a few authors could share a style.

5.4 Future Work

StyleChooser is based on lexical attraction, which uses word choice to find word dependencies in a sentence. Since authorship is not the only feature that affects word choice, this program could possibly be used to distinguish between appropriate audiences, genres or topics of texts. This opens up a whole new avenue of possible uses for my program. Any of these categories are very useful for computer agents to be able to distinguish between. If a search engine could have a profile of documents that were the right grade level for a searcher, for example, it could return results that have a similar audience (determined using a lexical attraction based language model) earlier in the results list, these being documents that are more likely to be useful to the searcher.

Another important aspect of StyleChooser is that it does not use qualities inherent in English at all, and should work for other languages. It would be interesting to see if varying the training and test set sizes have the same effect in other languages. Another interesting study would be to see if the slopes and intersects of the best fit lines for the average mutual information vs. training set length differ greatly between authors and between languages.

Another possible use for a lexical attraction based style differentiator is in a program to stylize computer generated text. StyleChooser gives a way to differentiate between a sentence in one style and a sentence in another, so if it could be paired up with a good representation for meaning it could be used to search for sentences in a particular style that match a particular meaning. This could lead to 'friendlier' agents and smoother,

more natural human/computer interaction, because computers could communicate in a familiar style.

Which leads me to another question: how well would StyleChooser work on transcribed dialogue? Do the differences between written text and spoken language cause StyleChooser to perform better or worse, or do they have no effect?

These questions all interest me because they take lexical attraction and try to confront some truth about language and the humans use it. I think lexical attraction could be a great tool for trying to further understand language and how humans think.

6. Contributions

In this thesis, I introduced lexical attraction as a tool for determining authorship. Lexical attraction is based on the interplay between words within sentences, which is directly related to word choice. The author of a text has a huge impact on word choice, so lexical attraction can be used to differentiate between authors.

I wrote a program, StyleChooser, to determine the authorship of texts based on statistical data about the dependencies of words within a sentence. StyleChooser is based on Yuret's lexical attraction parser, and uses the mutual information in the word dependencies determined by the author model as a metric for how well a particular author matches that text.

I tested and analyzed StyleChooser's performance and found that it had an accuracy of 75% on a large test set of 369 texts. The metric StyleChooser uses to determine authorship, mutual information/words in the test set/log(words in the training set), is independent of the size of the training and test sets. StyleChooser suffers very little degradation of accuracy when the test texts are restricted to only 15000 words, keeping an accuracy of 72%.

Appendix A: Texts from Project Gutenberg Used for Training and Testing

This appendix lists the texts that I downloaded from Project Gutenberg (http://promo.net/pg/) to use as input for testing StyleChooser.

Table 6: Texts from Project Gutenberg (Set A – one text per author)

Author	Title	Word Count
Alcott	Rose In Bloom	94116
Aldrich	The Story of a Bad Boy	56308
Alger	Driven From Home	52292
Allen	The Woman Who Did	45395
Altsheler	The Star of Gettysburg	94396
Appleton	Tom Swift and his Aerial Warship	43874
Austen	Northanger Abbey	76933
Bourget	Andre Cornelis	49865
Bower	The Flying U Ranch	41207
Brown	Memiors of Carwin, the Biloquist	23713
Buchan	The Moon Endureth	67486
Burnett	A Lady of Quality	83420
Butler	The Way of All Flesh	162216
Call	Power through Repose	40708
Carlyle	Early Kings of Norway	36931
Carroll	Sylvie and Bruno	64725
Cather	O Pioneers	56000
Chesterton	The Club of Queer Trades	44129
Collins	The Haunted Hotel	62291
Conrad	An Outcast of the Islands	104706
Cooper	The Monikins	140760
Darwin	Coral Reefs	85930
De Quincy	Confessions of an English Opium Eater	38657
Defoe	The Further Adventures of Robinson Crusoe	100190
Dickens	The Mystery of Edwin Drood	93835
Douglass	Narrative of the Life of Frederick Douglass	41240
Doyle	The Lost World	75558
Edgeworth	The Absentee	101182
Eliot	Silas Marner	70973
Ferber	Dawn O'Hara, the Girl who Laughed	60701
Fielding	Amelia	70365
Fitzgerald	Tales of the Jazz Age	85813
Forster	The Longest Journey	94000
Gaskell	Cranford	70601
Gissing	Eve's Ransom	54545
Hawthorne	Grandfather's Chair	56643
Henry	The Gentle Grafter	43661
Hichens	The Prophet of Berkeley Square	77908
Hume	The Green Mummy	88859
Irving	Astoria	164176

Author	Title	Word Count
Kipling	Stalky and Co	64274
Lang	The Lilac Fairy Book	98039
Lawrence	Sons and Lovers	160147
Leroux	The Phantom of the Opera	80511
Locke	The Fortunate Youth	98735
London	Adventure	70802
MacDonald	Lilith	93464
MacGrath	Man on the Box	67570
McCutcheon	The Hollow of her Hand	116562
Melville	Omoo	101149
Montgomery	Anne of Green Gables	102302
Norris	Harriet and the Piper	88363
Perkins	The French Twins	25646
Reade	A Simpleton	129950
Rinehart	The Man in Lower Ten	65185
Roe	He Fell in Love with his Wife	102580
Scott	Rob Roy	64185
Swift	Gulliver's Travels	103821
Thoreau	Walden	116598
Twain	Adventures of Tom Sawyer	71988
Van Dyke	The Blue Flower	51420
Wells	The War of the Worlds	60378

Table 7: Texts from Project Gutenberg (Set B – multiple texts per author)

Author	Title	Word Count
Alcott	Hospital Sketches	28840
Alcott	Flower Fables	34040
Alcott	A Modern Cinderella	40207
Alcott	Eight Cousins	72573
Alcott	Under the Lilacs	81772
Alcott	Jack and Jill	93517
Alcott	Jo's Boys	100638
Alcott	An Old-Fashioned Girl	103089
Alcott	Little Men	104843
Alcott	Work- A Story of Experience	118753
Alcott	Little Women	185665
Aldrich	Cruise of the Dolphin	4391
Aldrich	Majorie Daw	7401
Aldrich	An Old Town by the Sea	20301
Aldrich	Ponkapog Papers	29789
Alger	The Cash Boy	27587
Alger	Timothy Crump's Ward	40155
Alger	Phil, the Fiddler	42085
Alger	Paul the Peddler	42739
Alger	Joe the Hotel Boy	44152
Alger	The Errand Boy	53669
Alger	Cast Upon the Breakers	55949
Alger	Paul Prescott's Charge	59026
Alger	Frank's Campaign	61145
Allen	The British Barbarians	33652
Allen	Biographies of Working Men	47237
Allen	An African Millionaire	65371
Allen	Hilda Wade, A Woman With Tenacity Of Purpose	87985

Author	Title	Word Count
Allen	What's Bred in the Bone	98153
Altsheler	The Guns of Shiloh	86874
Altsheler	The Guns of Bull Run	89406
Altsheler	The Scouts of Stonewall	90131
Altsheler	The Scouts of the Vally	108124
Appleton	Tom Swift Among the Diamond Makers	41290
Appleton	Tom Swift and his Air Glider	41605
Appleton	Tom Swift in the Caves of Ice	41703
Appleton	Tow Swift Among the Fire Fighters	43636
Appleton	Tom Swift and his Undersea Search	45059
Appleton	Tom Swift and his Air Scout	45677
Austen	Lady Susan	23045
Austen	Persuasion	83309
Austen	Sense and Sensibility	118667
Austen	Pride and Prejudice	121498
Austen	Emma	158183
Austen	Mansfield Park	159646
Bourget	Cosmopolis	99959
Bower	Rowdy of the Cross L	23233
Bower	The Lure of the Dim Trails	28909
Bower	Her Prairie Knight	33202
Bower	The Trail of the White Mule	54436
Bower	Cabin Fever	56036
Bower	The Heritage of the Souix	58871
Bower	Cow Country	68692
Bower	Jean of the Lazy A	72272
Bower	The Flying U's Last Stand	78658
Bower	Good Indian	79425
Brown	Wieland	82258
Buchan	The Thirty-Nine Steps	41035
Buchan	Prester John	76315
Buchan	The Hunting Tower	78729
Buchan	The Path of the King	85671
Buchan	Greenmantle	98925
Buchan	Mr Standfast	129023
Burnett	In the Closed Room	11130
Burnett	The Dawn of a Tomorrow	18532
Burnett	The White People	21322
Burnett	Little Lord Fauntleroy	58322
Burnett	A Little Princess	66331
Burnett	The Secret Garden	80538
Burnett	The Lost Prince	99200
Burnett	The Shuttle	212707
Butler	The Fair Haven	81237
Call	As a Matter of Course	23264
Call	The Freedom of Life	31756
Call	Nerves and Common Sense	54743
Carlyle	Latter Day Pamphlets	17081
Carlyle	Life of John Sterling	87452
Carlyle	On Heroes and Hero Warship and the Heroic	88387
Carlyle	History of Friedrich the II of Prussia Volume 21	118685
Carlyle	The French Revolution	293958
Carroll	Alice in Wonderland	26455
Carroll	Through the Looking Glass	29284

Author	Title	Word Count
Chesterton	Manalive	57629
Chesterton	The Man Who Was Thursday	57931
Chesterton	The Man Who Knew Too Much	59668
Chesterton	The Wisdom of Father Brown	71616
Chesterton	The Innocence of Father Brown	78893
Collins	The Two Destinies	88937
Collins	The Black Robe	106706
Collins	The Evil Genius	109775
Collins	Basil	113839
Collins	I Say No	118193
Collins	The Legacy of Cain	118818
Collins	After Dark	135410
Collins	Antonia	165534
Collins	Man and Wife	226786
Collins	The Woman in White	244918
Collins	No Name	267387
Collins	Armadale	296302
Conrad	Heart of Darknes	38649
Conrad	The Mirror of the Sea	60864
Conrad	Tales of Unrest	61410
Conrad	Within the Tides	62524
Conrad	Twixt Land and Sea	71604
Conrad	A Set of Six	84564
Conrad	The Arrow of Gold	102553
Conrad	Chance	137516
Cooper	Imagination and Heart	39351
Cooper	Autobiography of a Pocket Handkerchief	53095
Cooper	The Last of the Mohicans	147159
Cooper	The Pioneers	170276
Cooper	Oak Openings	172163
Cooper	Jack Tier	179408
Cooper	The Deerslayer	225603
Darwin	The Autobiography of Charles Darwin	22434
Darwin	Volcanic Islands	57152
Darwin	The Expression of the Emotions in Man and Animals	107741
Darwin	Geological Observations On South America	134334
Darwin	Origin of Species	202507
Darwin	A Naturalist's Voyage Around the World	206388
Darwin	The Descent of Man	307650
De Quincy	The Ceasars	70628
Defoe	From London to Land's End	35245
Defoe	An Essay Upon Projects	46568
Defoe	Robinson Crusoe	121510
Defoe	The Fortunes and Misfortunes of the Famous Moll Flanders	136412
Dickens	Somebody's Luggage	19315
Dickens	A Christmas Carol	28285
Dickens	The Chimes	30959
Dickens	The Cricket on the Hearth	31850
Dickens	A House to Let	33907
Dickens	Hard Times	104146
Dickens	The Haunted House	130554
Dickens	A Tale of Two Cities	135702
Dickens	The Uncommercial Traveller	143914
Dickens	Oliver Twist	157116

Author	Title	Word Count
Dickens	A Child's History of England	163926
Dickens	Great Expectations	186261
Dickens	The Pickwick Papers	298280
Dickens	Nicholas Nickleby	320609
Dickens	MartinChuzzlewit	335527
Dickens	Bleak House	352738
Dickens	David Copperfield	358439
Dickens	Dombey and Son	358839
Douglass	Collected Articles of Frederick Douglass	8079
Douglass	My Bondage and My Freedom	127650
Doyle	The Disappearance of Lady Frances Carfax	7675
Doyle	The Vital Message	28586
Doyle	A Study in Scarlet	43550
Doyle	The Hound of the Baskervilles	59132
Doyle	The Adventures of Sherlock Holmes	104421
Doyle	The Return of Sherlock Holmes	112035
Doyle	The Great Boer War	223306
Edgeworth	Castle Rackrent	34481
Edgeworth	Murad the Unlucky	46803
Edgeworth	The Parent's Assistant	166057
Eliot	Brother Jacob	16632
Eliot	Adam Bede	213853
Eliot	Middlemarch	316660
Ferber	Emma McChesney and Co.	41725
Ferber	Buttered Side Down	44677
Ferber	One Basket	47917
Ferber	Fanny Herself	98576
Fielding	Journal of a Voyage to Lisbon	36129
Fielding	From this World to the Next	45785
Fielding	History of Tom Jones, A Foundling	345347
Fitzgerald	Flappers and Philosophers	60519
Fitzgerald	This Side of Paradise	80453
Forster	Where Angels Fear to Tread	49475
Forster	A Room with a View	66502
Forster	Howards End	108676
Gaskell	Half a Life-Time Ago	17724
Gaskell	Cousin Phillis	40097
Gaskell	A Dark Night's Work	67015
Gaskell	My Lady Ludlow	77168
Gaskell	Life of Charlotte Bronte	83164
Gaskell	Ruth	160415
Gaskell	Mary Barto	160456
Gaskell	North and South	181475
Gaskell	Sylvia's Lovers	190162
Gaskell	Wives and Daughters	268195
Gissing	By the Ionian Sea	39911
Gissing	Denzil Quarrier	79560
Gissing	The Crown of Life	118995
Gissing	In the Year of Jubilee	130458
Gissing	The Odd Women	138225
Gissing	A Life's Morning	141083
Gissing	The Emancipated	143157
Gissing	Born in Exile	157485
Gissing	The Nether World	160116

Author	Title	Word Count
Gissing	New Grub Street	185732
Gissing	Demos	194967
Hawthorne	The Great Stone Face	19069
Hawthorne	The Marble Faun Volume 1	62419
Hawthorne	Tanglewood Tales	65507
Hawthorne	The Blithedale Romance	75765
Hawthorne	The Marble Faun Volume 2	76834
Hawthorne	From Twice Told Storie	84793
Hawthorne	House of the Seven Gables	102365
Henry	Waifs and Strays Part 1	29011
Henry	The Four Million	51897
Henry	Voice of the City	54145
Henry	Strictly Business	70724
Hichens	The Spell of Egypt	36181
Hichens	A Spirit in Person	187515
Hichens	December Love	188235
Hichens	The Garden of Allah	196942
Hume	The Mystery of a Hansom Cab	86514
Hume	The Secret Passage	87691
Hume	Madame Midas	105083
Irving	The Legend of Sleepy Hollow	11811
Irving	Old Christmas	18381
Irving	Adventures of Captain Bonneville	116309
Irving	The Sketch-Book of Geoffrey Crayon	126035
Irving	Life of George Washington	127538
Irving	Chronicle of the Conquest of Grananda	156160
Kipling	Just So Stories	28728
Kipling	The Man Who Would Be King	42271
Kipling	The Jungle Book	50811
Kipling	Captains Courageous	53398
Kipling	Puck of Pook' Hill	58812
Kipling	The Second Jungle Book	63775
Kipling	Soldiers Three	65072
Kipling	Actions and Reactions	65914
Kipling	Plain Tales from the Hills	71846
Kipling	Rewards and Fairies	73894
Kipling	The Day's Work	106522
Kipling	Kim	107781
Kipling	Burning Daylight	112642
Lang	The Brown Fairy Book	96827
Lang	The Crimson Fairy Book	97221
Lang	The Orange Fairy Book	98019
Lang	The Violet Fairy Book	101867
Lang	A Monk of Fife	105057
Lang	The Yellow Fairy Book	110042
Lang	The Arabian Nights	110517
Lang	The Red Fairy Book	124431
Lang	The Blue Fairy Book	138364
Lawrence	Aaron's Rod	112195
Lawrence	Women in Love	180230
Leroux	Mystery ofthe Yellow Room	74766
Leroux	The Secret of the Night	97278
Locke	Simon the Jester	98683
Locke	The Red Planet	101238

Author	Title	Word Count
London	A Collection of Stories	22903
London	The House of Pride and Other Tales of Hawaii	30305
London	The Call of the Wild	31797
London	Before Adam	38989
London	Lost Face	40996
London	The Faith of Men	46615
London	Moon-Face and Other Stories	47939
London	The Son of the Wolf	48526
London	Love of Life and Other Stories	48840
London	The Night Born	52036
London	On the Makaloa Mat	55119
London	The People of the Abyss	61640
London	John Barleycorn	64694
London	Jerry of the Islands	69113
London	The Cruise ofthe Snark	79418
London	The Iron Heel	86602
London	Michael, Brother of Jerry	95394
london	Jacket	103673
London	The Sea Wolf	106279
London	The Mutiny of the Elsinore	113503
London	Martin Eden	140012
MacDonald	The Diary of an Old Soul	21648
MacDonald	The Princess and the Goblin	51340
MacDonald	The Princess and Curdie	56392
MacDonald	At the Back of the North Wind	88626
MacDonald	Donal Grant	175187
MacDonald	Sir Gibbie	182354
MacDonald	Robert Falconer	207728
MacGrath	Half a Rogue	86593
MacGrath	The Drums of Jeopardy	87104
MacGrath	The Puppet Crown	105760
McCutcheon	Brewster's Millions	62534
McCutcheon	Beverly of Graustark	85651
McCutcheon	Graustark	91054
McCutcheon	Viola Gwyn	107535
McCutcheon	The Rose in the Ring	124896
Melville	Typee	106778
Melville	Moby Dick	208425
Montgomery	The Golden Road	75525
Montgomery	Anne of the Island	77667
Montgomery	Anne's House of Dreams	79259
Montgomery	Anne of Avonlea	89502
Montgomery	Rilla of Ingleside	103320
Norris	Undertow	34770
Norris	The Rich Mrs. Brgoyne	42810
Norris	The Heart of Rachael	127489
Norris	The Story of Julia Page	136457
Norris	Saturday's Child	164577
Perkins	The Swiss Twins	17060
Perkins	The Japanese Twins	19265
Perkins	The Dutch Twins	21068
Perkins	The Eskimo Twins	22122
Perkins	The Belgian Twins	22971
Reade	Christie Johnstone	49374

Author	Title	Word Count
Reade	Peg Woffington	54071
Reade	White Lies	119746
Reade	Foul Play	155178
Reade	A Woman-Hater	156039
Reade	Put Yourself in His Place	204254
Reade	Hard Cash	261479
Rinehart	The Confession	29313
Rinehart	Sight Unseen	34520
Rinehart	The After House	48536
Rinehart	When a Man Marries	54923
Rinehart	The Bat	63864
Rinehart	Where there's a Will	65405
Rinehart	The Circular Staircase	69826
Rinehart	The Amazing Interlude	71167
Rinehart	The Street of Seven Stars	75807
Rinehart	Bab - a Sub-Deb	84942
Rinehart	Tish	88227
Rinehart	Long Live thy King	111794
Rinehart	The Breaking Point	114228
Rinehart	Dangerous Days	120182
Rinehart	A Poor Wise Man	124993
Roe	A Day of Fate	108436
Roe	Success with Small Fruits	108492
Roe	What Can She Do	123132
Roe	From Jest to Earnest	128325
Roe	The Earth Trembled	136894
Roe	Barriers Burned Away	141649
Roe	A Face Illumined	166743
Scott	The Black Dwarf	57495
Scott	Chronicles of the Canongate	72322
Scott	The Antiquary	78980
Scott	The Lady of the Lake	80655
Scott	A Legend of Montrose	82030
Scott	Waverly	100752
Scott	Bride of Lammermoor	122728
Scott	The Talisman	124788
Scott	Redgauntlet	180449
Scott	Kenilworth	184207
Scott	Ivanhoe	192211
Swift	Battle of the Books	38608
Swift	A Tale of a Tub	38979
Swift	The Journal to Stella	229490
Thoreau	Walking	12104
Thoreau	A Week on the Concord and Merrimack Rivers	115114
Twain	A Horse's Tale	17317
Twain	A Double Barrelled Detective Story	19180
Twain	Those Extraordinary Twins	19802
Twain	The American Claimant	62272
Twain	Adventures of Huckleberry Finn	110211
Twain	A Connecticut Yankee in King Arthur's Court	115103
Twain	Life on the Mississippi	144188
Twain	A Tramp Abroad	153418
Twain	Roughing It	163981
Twain	Following the Equator	185135

Author	Title	Word Count
Van Dyke	Fisherman's Luck and Some Other Uncertain Things	46587
Van Dyke	The Ruling Passion	54162
Wells	The Time Machine	32337
Wells	God the Invisible King	37440
Wells	The Island of Doctor Moreau	43411
Wells	First and Last Things	55289
Wells	The World Set Free	63932
Wells	Secret Places of the Heart	64791
Wells	The First Men in the Moon	69198
Wells	Twelve Stories and a Dream	73081
Wells	Soul of a Bishop	78922
Wells	In the Days of the Comet	81377
Wells	When the Sleeper Wakes	83429
Wells	Mankind in the Making	102109
Wells	The Research Magnificent	112916
Wells	Tono Bungay	133642
Wells	The New Machiavelli	145728

Appendix B: Full Results from Tests with StyleChooser

Many of the tests I performed with StyleChooser involved a result for each of 350 to 400 texts. In this Appendix I list out results for my tests.

Table 8: Ranking of correct author for each text after training on Set A, with original metric of MI/words in test set

A Child's History of England: correct author Dickens ranked 14
A Christmas Carol: correct author Dickens ranked 3
A Collection of Stories: correct author London ranked 5
A Connecticut Yankee in King Arthur's Court: correct author Twain ranked 10
A Dark Night's Work: correct author Gaskell ranked 5
A Day of Fate: correct author Roe ranked 1
A Double Barrelled Detective Story: correct author Twain ranked 6
A Face Illumined: correct author Roe ranked 1
A Horse's Tale: correct author Twain ranked 7
A House to Let: correct author Dickens ranked 4
A Legend of Montrose: correct author Scott ranked 5
A Life's Morning: correct author Gissing ranked 6
A Little Princess: correct author Burnett ranked 20
A Modern Cinderella: correct author Alcott ranked 1
A Monk of Fife: correct author Lang ranked 1
A Naturalist's Voyage Around the World: correct author Darwin ranked 4
A Poor Wise Man: correct author Rinehart ranked 6
A Room with a View: correct author Forster ranked 1
A Set of Six: correct author Conrad ranked 1
A Spirit in Person: correct author Hichens ranked 12
A Study in Scarlet: correct author Doyle ranked 4
A Tale of a Tub: correct author Swift ranked 1
A Tale of Two Cities: correct author Dickens ranked 1
A Tramp Abroad: correct author Twain ranked 10
A Week on the Concord and Merrimack Rivers: correct author Thoreau ranked 1
A Woman-Hater: correct author Reade ranked 1
Aaron's Rod: correct author Lawrence ranked 28
Actions and Reactions: correct author Kipling ranked 17
Adam Bede: correct author Eliot ranked 1
Adventures of Captain Bonneville: correct author Irving ranked 1
Adventures of Huckleberry Finn: correct author Twain ranked 1
After Dark: correct author Collins ranked 11
Alice in Wonderland: correct author Carroll ranked 1
An African Millionaire: correct author Allen ranked 40
An Essay Upon Projects: correct author Defoe ranked 8
An Old Town by the Sea: correct author Aldrich ranked 1
An Old-Fashioned Girl: correct author Alcott ranked 1
Anne of Avonlea: correct author Montgomery ranked 1
Anne of the Island: correct author Montgomery ranked 1

Anne's House of Dreams: correct author Montgomery ranked 1
Antonia: correct author Collins ranked 35
Armadale: correct author Collins ranked 4
As a Matter of Course: correct author Call ranked 1
At the Back of the North Wind: correct author MacDonald ranked 10
Autobiography of a Pocket Handkerchief: correct author Cooper ranked 1
Bab - a Sub-Deb: correct author Rinehart ranked 6
Barriers Burned Away: correct author Roe ranked 2
Basil: correct author Collins ranked 8
Battle of the Books: correct author Swift ranked 1
Before Adam: correct author London ranked 5
Beverly of Graustark: correct author McCutcheon ranked 1
Biographies of Working Men: correct author Allen ranked 45
Bleak House: correct author Dickens ranked 1
Born in Exile: correct author Gissing ranked 4
Brewster's Millions: correct author McCutcheon ranked 1
Bride of Lammermoor: correct author Scott ranked 1
Brother Jacob: correct author Eliot ranked 2
Burning Daylight: correct author Kipling ranked 34
Buttered Side Down: correct author Ferber ranked 1
By the Ionian Sea: correct author Gissing ranked 14
Cabin Fever: correct author Bower ranked 21
Captains Courageous: correct author Kipling ranked 5
Cast Upon the Breakers: correct author Alger ranked 1
Castle Rackrent: correct author Edgeworth ranked 1
Chance: correct author Conrad ranked 1
Christie Johnstone: correct author Reade ranked 1
Chronicle of the Conquest of Grananda: correct author Irving ranked 1
Chronicles of the Canongate: correct author Scott ranked 5
Collected Articles of Frederick Douglass: correct author Douglass ranked 3
Cosmopolis: correct author Bourget ranked 9
Cousin Phillis: correct author Gaskell ranked 1
Cow Country: correct author Bower ranked 12
Cruise of the Dolphin: correct author Aldrich ranked 1
Dangerous Days: correct author Rinehart ranked 6
David Copperfield: correct author Dickens ranked 2
December Love: correct author Hichens ranked 10
Demos: correct author Gissing ranked 8
Denzil Quarrier: correct author Gissing ranked 5
Dombey and Son: correct author Dickens ranked 1
Donal Grant: correct author MacDonald ranked 7
Eight Cousins: correct author Alcott ranked 1
Emma McChesney and Co.: correct author Ferber ranked 6
Emma: correct author Austen ranked 1
Fanny Herself: correct author Ferber ranked 6
First and Last Things: correct author Wells ranked 26
Fisherman's Luck and Some Other Uncertain Things: correct author Van Dyke ranked 20
Flappers and Philosophers: correct author Fitzgerald ranked 1
Flower Fables: correct author Alcott ranked 7
Following the Equator: correct author Twain ranked 24
Foul Play: correct author Reade ranked 1
Frank's Campaign: correct author Alger ranked 8
From Jest to Earnest: correct author Roe ranked 1
From London to Land's End: correct author Defoe ranked 3

From this World to the Next: correct author Fielding ranked 2
From Twice Told Stories[Hawthorne ranked 1
Geological Observations On South America: correct author Darwin ranked 1
God the Invisible King: correct author Wells ranked 23
Good Indian: correct author Bower ranked 13
Graustark: correct author McCutcheon ranked 1
Great Expectations: correct author Dickens ranked 1
Greenmantle: correct author Buchan ranked 1
Half a Life-Time Ago: correct author Gaskell ranked 11
Half a Rogue: correct author MacGrath ranked 3
Hard Cash: correct author Reade ranked 1
Hard Times: correct author Dickens ranked 1
Heart of Darknes: correct author Conrad ranked 1
Her Prairie Knight: correct author Bower ranked 25
Hilda Wade, A Woman With Tenacity Of Purpose: correct author Allen ranked 29
History of Friedrich the II of Prussia Volume 21: correct author Carlyle ranked 34
History of Tom Jones, A Foundling: correct author Fielding ranked 1
Hospital Sketches: correct author Alcott ranked 1
House of the Seven Gables: correct author Hawthorne ranked 14
Howards End: correct author Forster ranked 1
I Say No: correct author Collins ranked 1
Imagination and Heart: correct author Cooper ranked 1
In the Closed Room: correct author Burnett ranked 16
In the Days of the Comet: correct author Wells ranked 3
In the Year of Jubilee: correct author Gissing ranked 4
Ivanhoe: correct author Scott ranked 4
Jack and Jill: correct author Alcott ranked 1
Jack Tier: correct author Cooper ranked 1
Jacket: correct author london ranked 7
Jean of the Lazy A: correct author Bower ranked 26
Jerry of the Islands: correct author London ranked 1
Jo's Boys: correct author Alcott ranked 1
Joe the Hotel Boy: correct author Alger ranked 6
John Barleycorn: correct author London ranked 4
Journal of a Voyage to Lisbon: correct author Fielding ranked 4
Just So Stories: correct author Kipling ranked 14
Kenilworth: correct author Scott ranked 4
Kim: correct author Kipling ranked 13
Lady Susan: correct author Austen ranked 1
Latter Day Pamphlets: correct author Carlyle ranked 14
Life of Charlotte Bronte: correct author Gaskell ranked 10
Life of George Washington: correct author Irving ranked 1
Life of John Sterling: correct author Carlyle ranked 41
Life on the Mississippi: correct author Twain ranked 11
Little Lord Fauntleroy: correct author Burnett ranked 32
Little Men: correct author Alcott ranked 1
Little Women: correct author Alcott ranked 1
Long Live thy King: correct author Rinehart ranked 12
Lost Face: correct author London ranked 1
Love of Life and Other Stories: correct author London ranked 3
Madame Midas: correct author Hume ranked 2
Majorie Daw: correct author Aldrich ranked 2
Man and Wife: correct author Collins ranked 2
Manalive: correct author Chesterton ranked 18

Mankind in the Making: correct author Wells ranked 22
Mansfield Park: correct author Austen ranked 1
Martin Eden: correct author London ranked 3
MartinChuzzlewit: correct author Dickens ranked 1
Mary Barton[Gaskell ranked 5
Michael, Brother of Jerry: correct author London ranked 1
Middlemarch: correct author Eliot ranked 6
Moby Dick: correct author Melville ranked 1
Moon-Face and Other Stories: correct author London ranked 4
Mr Standfast: correct author Buchan ranked 1
Murad the Unlucky: correct author Edgeworth ranked 1
My Bondage and My Freedom: correct author Douglass ranked 1
My Lady Ludlow: correct author Gaskell ranked 2
Mystery ofthe Yellow Room: correct author Leroux ranked 1
Nerves and Common Sense: correct author Call ranked 2
New Grub Street: correct author Gissing ranked 4
Nicholas Nickleby: correct author Dickens ranked 3
No Name: correct author Collins ranked 1
North and South: correct author Gaskell ranked 6
Oak Openings: correct author Cooper ranked 2
Old Christmas: correct author Irving ranked 2
Oliver Twist: correct author Dickens ranked 3
On Heroes and Hero Warship and the Heroic: correct author Carlyle ranked 21
On the Makaloa Mat: correct author London ranked 6
One Basket: correct author Ferber ranked 6
Origin of Species: correct author Darwin ranked 1
Paul Prescott's Charge: correct author Alger ranked 6
Paul the Peddler: correct author Alger ranked 3
Peg Woffington: correct author Reade ranked 1
Persuasion: correct author Austen ranked 1
Phil, the Fiddler: correct author Alger ranked 4
Plain Tales from the Hills: correct author Kipling ranked 21
Ponkapog Papers: correct author Aldrich ranked 8
Prester John: correct author Buchan ranked 1
Pride and Prejudice: correct author Austen ranked 1
Puck of Pook' Hill: correct author Kipling ranked 23
Put Yourself in His Place: correct author Reade ranked 1
Redgauntlet: correct author Scott ranked 2
Rewards and Fairies: correct author Kipling ranked 15
Rilla of Ingleside: correct author Montgomery ranked 1
Robert Falconer: correct author MacDonald ranked 12
Robinson Crusoe: correct author Defoe ranked 1
Roughing It: correct author Twain ranked 9
Rowdy of the Cross L: correct author Bower ranked 5
Ruth: correct author Gaskell ranked 6
Saturday's Child: correct author Norris ranked 1
Secret Places of the Heart: correct author Wells ranked 9
Sense and Sensibility: correct author Austen ranked 1
Sight Unseen: correct author Rinehart ranked 1
Simon the Jester: correct author Locke ranked 1
Sir Gibbie: correct author MacDonald ranked 6
Soldiers Three: correct author Kipling ranked 1
Somebody's Luggage: correct author Dickens ranked 2
Soul of a Bishop: correct author Wells ranked 14

Strictly Business: correct author Henry ranked 18
Success with Small Fruits: correct author Roe ranked 7
Sylvia's Lovers: correct author Gaskell ranked 5
Tales of Unrest: correct author Conrad ranked 1
Tanglewood Tales: correct author Hawthorne ranked 8
The Adventures of Sherlock Holmes: correct author Doyle ranked 3
The After House: correct author Rinehart ranked 2
The Amazing Interlude: correct author Rinehart ranked 9
The American Claimant: correct author Twain ranked 7
The Antiquary: correct author Scott ranked 1
The Arabian Nights: correct author Lang ranked 1
The Arrow of Gold: correct author Conrad ranked 1
The Autobiography of Charles Darwin: correct author Darwin ranked 14
The Bat: correct author Rinehart ranked 2
The Belgian Twins: correct author Perkins ranked 8
The Black Dwarf: correct author Scott ranked 1
The Black Robe: correct author Collins ranked 2
The Blithedale Romance: correct author Hawthorne ranked 19
The Blue Fairy Book: correct author Lang ranked 1
The Breaking Point: correct author Rinehart ranked 6
The British Barbarians: correct author Allen ranked 7
The Brown Fairy Book: correct author Lang ranked 1
The Call of the Wild: correct author London ranked 3
The Cash Boy: correct author Alger ranked 2
The Ceasars: correct author De Quincy ranked 15
The Chimes: correct author Dickens ranked 3
The Circular Staircase: correct author Rinehart ranked 1
The Confession: correct author Rinehart ranked 1
The Cricket on the Hearth: correct author Dickens ranked 1
The Crimson Fairy Book: correct author Lang ranked 1
The Crown of Life: correct author Gissing ranked 4
The Cruise ofthe Snark: correct author London ranked 1
The Dawn of a Tomorrow: correct author Burnett ranked 15
The Day's Work: correct author Kipling ranked 6
The Deerslayer: correct author Cooper ranked 1
The Descent of Man: correct author Darwin ranked 4
The Diary of an Old Soul: correct author MacDonald ranked 1
The Disappearance of Lady Frances Carfax: correct author Doyle ranked 1
The Drums of Jeopardy: correct author MacGrath ranked 6
The Dutch Twins: correct author Perkins ranked 30
The Earth Trembled: correct author Roe ranked 1
The Emancipated: correct author Gissing ranked 5
The Errand Boy: correct author Alger ranked 1
The Eskimo Twins: correct author Perkins ranked 38
The Evil Genius: correct author Collins ranked 1
The Expression of the Emotions in Man and Animals: correct author Darwin ranked 7
The Fair Haven: correct author Butler ranked 1
The Faith of Men: correct author London ranked 4
The First Men in the Moon: correct author Wells ranked 1
The Flying U's Last Stand: correct author Bower ranked 1
The Fortunes and Misfortunes of the Famous Moll Flanders: correct author Defoe ranked 1
The Four Million: correct author Henry ranked 31
The Freedom of Life: correct author Call ranked 2
The French Revolution: correct author Carlyle ranked 38

The Garden of Allah: correct author Hichens ranked 12
The Golden Road: correct author Montgomery ranked 1
The Great Boer War: correct author Doyle ranked 7
The Great Stone Face: correct author Hawthorne ranked 8
The Guns of Bull Run: correct author Altsheler ranked 1
The Guns of Shiloh: correct author Altsheler ranked 1
The Haunted House: correct author Dickens ranked 10
The Heart of Rachael: correct author Norris ranked 1
The Heritage of the Souix: correct author Bower ranked 4
The Hound of the Baskervilles: correct author Doyle ranked 1
The House of Pride and Other Tales of Hawaii: correct author London ranked 1
The Hunting Tower: correct author Buchan ranked 3
The Innocence of Father Brown: correct author Chesterton ranked 14
The Iron Heel: correct author London ranked 6
The Island of Doctor Moreau: correct author Wells ranked 1
The Japanese Twins: correct author Perkins ranked 28
The Journal to Stella: correct author Swift ranked 5
The Jungle Book: correct author Kipling ranked 12
The Lady of the Lake: correct author Scott ranked 8
The Last of the Mohicans: correct author Cooper ranked 2
The Legacy of Cain: correct author Collins ranked 2
The Legend of Sleepy Hollow: correct author Irving ranked 1
The Lost Prince: correct author Burnett ranked 16
The Lure of the Dim Trails: correct author Bower ranked 11
The Man Who Knew Too Much: correct author Chesterton ranked 19
The Man Who Was Thursday: correct author Chesterton ranked 7
The Man Who Would Be King: correct author Kipling ranked 37
The Marble Faun Volume 1: correct author Hawthorne ranked 14
The Marble Faun Volume 2: correct author Hawthorne ranked 11
The Mirror of the Sea: correct author Conrad ranked 1
The Mutiny of the Elsinore: correct author London ranked 1
The Mystery of a Hansom Cab: correct author Hume ranked 1
The Nether World: correct author Gissing ranked 5
The New Machiavelli: correct author Wells ranked 5
The Night Born: correct author London ranked 1
The Odd Women: correct author Gissing ranked 4
The Orange Fairy Book: correct author Lang ranked 1
The Parent's Assistant: correct author Edgeworth ranked 1
The Path of the King: correct author Buchan ranked 2
The People of the Abyss: correct author London ranked 14
The Pickwick Papers: correct author Dickens ranked 3
The Pioneers: correct author Cooper ranked 2
The Princess and Curdie: correct author MacDonald ranked 3
The Princess and the Goblin: correct author MacDonald ranked 5
The Puppet Crown: correct author MacGrath ranked 1
The Red Fairy Book: correct author Lang ranked 1
The Red Planet: correct author Locke ranked 1
The Research Magnificent: correct author Wells ranked 7
The Return of Sherlock Holmes: correct author Doyle ranked 1
The Rich Mrs. Brgoyne: correct author Norris ranked 1
The Rose in the Ring: correct author McCutcheon ranked 1
The Ruling Passion: correct author Van Dyke ranked 15
The Scouts of Stonewall: correct author Altsheler ranked 1
The Scouts of the Vally: correct author Altsheler ranked 1

The Sea Wolf: correct author London ranked 1
The Second Jungle Book: correct author Kipling ranked 14
The Secret Garden: correct author Burnett ranked 32
The Secret of the Night: correct author Leroux ranked 1
The Secret Passage: correct author Hume ranked 1
The Shuttle: correct author Burnett ranked 10
The Sketch-Book of Geoffrey Crayon: correct author Irving ranked 1
The Son of the Wolf: correct author London ranked 9
The Spell of Egypt: correct author Hichens ranked 16
The Story of Julia Page: correct author Norris ranked 1
The Street of Seven Stars: correct author Rinehart ranked 6
The Swiss Twins: correct author Perkins ranked 25
The Talisman: correct author Scott ranked 5
The Thirty-Nine Steps: correct author Buchan ranked 1
The Time Machine: correct author Wells ranked 1
The Trail of the White Mule: correct author Bower ranked 9
The Two Destinies: correct author Collins ranked 1
The Uncommercial Traveller: correct author Dickens ranked 2
The Violet Fairy Book: correct author Lang ranked 1
The Vital Message: correct author Doyle ranked 4
The White People: correct author Burnett ranked 22
The Wisdom of Father Brown: correct author Chesterton ranked 19
The Woman in White: correct author Collins ranked 3
The World Set Free: correct author Wells ranked 5
The Yellow Fairy Book: correct author Lang ranked 1
This Side of Paradise: correct author Fitzgerald ranked 1
Those Extraordinary Twins: correct author Twain ranked 4
Through the Looking Glass: correct author Carroll ranked 1
Timothy Crump's Ward: correct author Alger ranked 11
Tish: correct author Rinehart ranked 2
Tom Swift Among the Diamond Makers: correct author Appleton ranked 1
Tom Swift and his Air Glider: correct author Appleton ranked 1
Tom Swift and his Air Scout: correct author Appleton ranked 1
Tom Swift and his Undersea Search: correct author Appleton ranked 1
Tom Swift in the Caves of Ice: correct author Appleton ranked 1
Tono Bungay: correct author Wells ranked 5
Tow Swift Among the Fire Fighters: correct author Appleton ranked 1
Twelve Stories and a Dream: correct author Wells ranked 3
Twixt Land and Sea: correct author Conrad ranked 1
Typee: correct author Melville ranked 1
Under the Lilacs: correct author Alcott ranked 1
Undertow: correct author Norris ranked 1
Viola Gwyn: correct author McCutcheon ranked 1
Voice of the City: correct author Henry ranked 30
Volcanic Islands: correct author Darwin ranked 1
Waifs and Strays Part 1: correct author Henry ranked 34
Walking: correct author Thoreau ranked 1
Waverly: correct author Scott ranked 4
What Can She Do: correct author Roe ranked 1
What's Bred in the Bone: correct author Allen ranked 19
When a Man Marries: correct author Rinehart ranked 1
When the Sleeper Wakes: correct author Wells ranked 1
Where Angels Fear to Tread: correct author Forster ranked 1
Where there's a Will: correct author Rinehart ranked 1

White Lies: correct author Reade ranked 1
Wieland: correct author Brown ranked 7
Within the Tides: correct author Conrad ranked 1
Wives and Daughters: correct author Gaskell ranked 4
Women in Love: correct author Lawrence ranked 8
Work- A Story of Experience: correct author Alcott ranked 1

Table 9: Ranking of correct author for each text after training on Set A, with new metric of MI/words in test set/log(words in training set)

A Child's History of England: correct author Dickens ranked 20
A Christmas Carol: correct author Dickens ranked 1
A Collection of Stories: correct author London ranked 1
A Connecticut Yankee in King Arthur's Court: correct author Twain ranked 3
A Dark Night's Work: correct author Gaskell ranked 1
A Day ofFate: correct author Roe ranked 1
A Double Barrelled Detective Story: correct author Twain ranked 2
A Face Illumined: correct author Roe ranked 1
A Horse's Tale: correct author Twain ranked 1
A House to Let: correct author Dickens ranked 4
A Legend of Montrose: correct author Scott ranked 1
A Life's Morning: correct author Gissing ranked 1
A Little Princess: correct author Burnett ranked 24
A Modern Cinderella: correct author Alcott ranked 1
A Monk of Fife: correct author Lang ranked 3
A Naturalist's Voyage Around the World: correct author Darwin ranked 2
A Poor Wise Man: correct author Rinehart ranked 3
A Room with a View: correct author Forster ranked 1
A Set of Six: correct author Conrad ranked 1
A Spirit in Person: correct author Hichens ranked 9
A Study in Scarlet: correct author Doyle ranked 1
A Tale of a Tub: correct author Swift ranked 1
A Tale of Two Cities: correct author Dickens ranked 1
A Tramp Abroad: correct author Twain ranked 2
A Week on the Concord and Merrimack Rivers: correct author Thoreau ranked 1
A Woman-Hater: correct author Reade ranked 1
Aaron's Rod: correct author Lawrence ranked 52
Actions and Reactions: correct author Kipling ranked 5
Adam Bede: correct author Eliot ranked 1
Adventures of Captain Bonneville: correct author Irving ranked 1
Adventures of Huckleberry Finn: correct author Twain ranked 1
After Dark: correct author Collins ranked 1
Alice in Wonderland: correct author Carroll ranked 1
An African Millionaire: correct author Allen ranked 15
An Essay Upon Projects: correct author Defoe ranked 8
An Old Town by the Sea: correct author Aldrich ranked 1
An Old-Fashioned Girl: correct author Alcott ranked 1
Anne of Avonlea: correct author Montgomery ranked 1
Anne of the Island: correct author Montgomery ranked 1
Anne's House of Dreams: correct author Montgomery ranked 1
Antonia: correct author Collins ranked 17
Armadale: correct author Collins ranked 1
As a Matter of Course: correct author Call ranked 1
At the Back of the North Wind: correct author MacDonald ranked 15

Autobiography of a Pocket Handkerchief: correct author Cooper ranked 1
Bab - a Sub-Deb: correct author Rinehart ranked 1
Barriers Burned Away: correct author Roe ranked 1
Basil: correct author Collins ranked 1
Battle of the Books: correct author Swift ranked 1
Before Adam: correct author London ranked 3
Beverly of Graustark: correct author McCutcheon ranked 1
Biographies of Working Men: correct author Allen ranked 23
Bleak House: correct author Dickens ranked 1
Born in Exile: correct author Gissing ranked 1
Brewster's Millions: correct author McCutcheon ranked 1
Bride of Lammermoor: correct author Scott ranked 1
Brother Jacob: correct author Eliot ranked 1
Burning Daylight: correct author Kipling ranked 25
Buttered Side Down: correct author Ferber ranked 1
By the Ionian Sea: correct author Gissing ranked 1
Cabin Fever: correct author Bower ranked 2
Captains Courageous: correct author Kipling ranked 1
Cast Upon the Breakers: correct author Alger ranked 1
Castle Rackrent: correct author Edgeworth ranked 1
Chance: correct author Conrad ranked 1
Christie Johnstone: correct author Reade ranked 1
Chronicle of the Conquest of Grananda: correct author Irving ranked 1
Chronicles of the Canongate: correct author Scott ranked 1
Collected Articles of Frederick Douglass: correct author Douglass ranked 1
Cosmopolis: correct author Bourget ranked 1
Cousin Phillis: correct author Gaskell ranked 1
Cow Country: correct author Bower ranked 1
Cruise of the Dolphin: correct author Aldrich ranked 1
Dangerous Days: correct author Rinehart ranked 4
David Copperfield: correct author Dickens ranked 2
December Love: correct author Hichens ranked 6
Demos: correct author Gissing ranked 1
Denzil Quarrier: correct author Gissing ranked 1
Dombey and Son: correct author Dickens ranked 1
Donal Grant: correct author MacDonald ranked 5
Eight Cousins: correct author Alcott ranked 1
Emma McChesney and Co.: correct author Ferber ranked 2
Emma: correct author Austen ranked 1
Fanny Herself: correct author Ferber ranked 2
First and Last Things: correct author Wells ranked 16
Fisherman's Luck and Some Other Uncertain Things: correct author Van Dyke ranked 2
Flappers and Philosophers: correct author Fitzgerald ranked 1
Flower Fables: correct author Alcott ranked 5
Following the Equator: correct author Twain ranked 15
Foul Play: correct author Reade ranked 1
Frank's Campaign: correct author Alger ranked 1
From Jest to Earnest: correct author Roe ranked 1
From London to Land's End: correct author Defoe ranked 2
From this World to the Next: correct author Fielding ranked 1
From Twice Told Stories[Hawthorne ranked 1
Geological Observations On South America: correct author Darwin ranked 1
God the Invisible King: correct author Wells ranked 11
Good Indian: correct author Bower ranked 1
Graustark: correct author McCutcheon ranked 1

Great Expectations: correct author Dickens ranked 2
Greenmantle: correct author Buchan ranked 1
Half a Life-Time Ago: correct author Gaskell ranked 1
Half a Rogue: correct author MacGrath ranked 1
Hard Cash: correct author Reade ranked 1
Hard Times: correct author Dickens ranked 1
Heart of Darknes: correct author Conrad ranked 1
Her Prairie Knight: correct author Bower ranked 4
Hilda Wade, A Woman With Tenacity Of Purpose: correct author Allen ranked 1
History of Friedrich the II of Prussia Volume 21: correct author Carlyle ranked 1
History of Tom Jones, A Foundling: correct author Fielding ranked 1
Hospital Sketches: correct author Alcott ranked 1
House of the Seven Gables: correct author Hawthorne ranked 1
Howards End: correct author Forster ranked 1
I Say No: correct author Collins ranked 1
Imagination and Heart: correct author Cooper ranked 2
In the Closed Room: correct author Burnett ranked 21
In the Days of the Comet: correct author Wells ranked 1
In the Year of Jubilee: correct author Gissing ranked 1
Ivanhoe: correct author Scott ranked 1
Jack and Jill: correct author Alcott ranked 1
Jack Tier: correct author Cooper ranked 1
Jacket: correct author london ranked 2
Jean of the Lazy A: correct author Bower ranked 4
Jerry of the Islands: correct author London ranked 1
Jo's Boys: correct author Alcott ranked 1
Joe the Hotel Boy: correct author Alger ranked 1
John Barleycorn: correct author London ranked 2
Journal of a Voyage to Lisbon: correct author Fielding ranked 1
Just So Stories: correct author Kipling ranked 3
Kenilworth: correct author Scott ranked 1
Kim: correct author Kipling ranked 4
Lady Susan: correct author Austen ranked 1
Latter Day Pamphlets: correct author Carlyle ranked 1
Life of Charlotte Bronte: correct author Gaskell ranked 2
Life of George Washington: correct author Irving ranked 1
Life of John Sterling: correct author Carlyle ranked 6
Life on the Mississippi: correct author Twain ranked 2
Little Lord Fauntleroy: correct author Burnett ranked 38
Little Men: correct author Alcott ranked 1
Little Women: correct author Alcott ranked 1
Long Live thy King: correct author Rinehart ranked 2
Lost Face: correct author London ranked 1
Love of Life and Other Stories: correct author London ranked 1
Madame Midas: correct author Hume ranked 1
Majorie Daw: correct author Aldrich ranked 1
Man and Wife: correct author Collins ranked 1
Manalive: correct author Chesterton ranked 1
Mankind in the Making: correct author Wells ranked 7
Mansfield Park: correct author Austen ranked 1
Martin Eden: correct author London ranked 1
MartinChuzzlewit: correct author Dickens ranked 1
Mary Barton[Gaskell ranked 1
Michael, Brother of Jerry: correct author London ranked 1
Middlemarch: correct author Eliot ranked 1

Moby Dick: correct author Melville ranked 1
Moon-Face and Other Stories: correct author London ranked 1
Mr Standfast: correct author Buchan ranked 1
Murad the Unlucky: correct author Edgeworth ranked 1
My Bondage and My Freedom: correct author Douglass ranked 1
My Lady Ludlow: correct author Gaskell ranked 1
Mystery ofthe Yellow Room: correct author Leroux ranked 1
Nerves and Common Sense: correct author Call ranked 1
New Grub Street: correct author Gissing ranked 1
Nicholas Nickleby: correct author Dickens ranked 1
No Name: correct author Collins ranked 1
North and South: correct author Gaskell ranked 1
Oak Openings: correct author Cooper ranked 1
Old Christmas: correct author Irving ranked 17
Oliver Twist: correct author Dickens ranked 1
On Heroes and Hero Warship and the Heroic: correct author Carlyle ranked 1
On the Makaloa Mat: correct author London ranked 1
One Basket: correct author Ferber ranked 2
Origin of Species: correct author Darwin ranked 1
Paul Prescott's Charge: correct author Alger ranked 1
Paul the Peddler: correct author Alger ranked 1
Peg Woffington: correct author Reade ranked 1
Persuasion: correct author Austen ranked 1
Phil, the Fiddler: correct author Alger ranked 1
Plain Tales from the Hills: correct author Kipling ranked 8
Ponkapog Papers: correct author Aldrich ranked 1
Prester John: correct author Buchan ranked 1
Pride and Prejudice: correct author Austen ranked 1
Puck of Pook' Hill: correct author Kipling ranked 8
Put Yourself in His Place: correct author Reade ranked 1
Redgauntlet: correct author Scott ranked 1
Rewards and Fairies: correct author Kipling ranked 2
Rilla of Ingleside: correct author Montgomery ranked 1
Robert Falconer: correct author MacDonald ranked 18
Robinson Crusoe: correct author Defoe ranked 1
Roughing It: correct author Twain ranked 3
Rowdy of the Cross L: correct author Bower ranked 1
Ruth: correct author Gaskell ranked 1
Saturday's Child: correct author Norris ranked 1
Secret Places of the Heart: correct author Wells ranked 2
Sense and Sensibility: correct author Austen ranked 1
Sight Unseen: correct author Rinehart ranked 1
Simon the Jester: correct author Locke ranked 1
Sir Gibbie: correct author MacDonald ranked 5
Soldiers Three: correct author Kipling ranked 1
Somebody's Luggage: correct author Dickens ranked 1
Soul of a Bishop: correct author Wells ranked 3
Strictly Business: correct author Henry ranked 1
Success with Small Fruits: correct author Roe ranked 9
Sylvia's Lovers: correct author Gaskell ranked 1
Tales of Unrest: correct author Conrad ranked 1
Tanglewood Tales: correct author Hawthorne ranked 2
The Adventures of Sherlock Holmes: correct author Doyle ranked 1
The After House: correct author Rinehart ranked 1
The Amazing Interlude: correct author Rinehart ranked 3

The American Claimant: correct author Twain ranked 2
The Antiquary: correct author Scott ranked 1
The Arabian Nights: correct author Lang ranked 1
The Arrow of Gold: correct author Conrad ranked 1
The Autobiography of Charles Darwin: correct author Darwin ranked 16
The Bat: correct author Rinehart ranked 1
The Belgian Twins: correct author Perkins ranked 1
The Black Dwarf: correct author Scott ranked 1
The Black Robe: correct author Collins ranked 1
The Blithedale Romance: correct author Hawthorne ranked 2
The Blue Fairy Book: correct author Lang ranked 1
The Breaking Point: correct author Rinehart ranked 1
The British Barbarians: correct author Allen ranked 1
The Brown Fairy Book: correct author Lang ranked 1
The Call of the Wild: correct author London ranked 1
The Cash Boy: correct author Alger ranked 1
The Ceasars: correct author De Quincy ranked 1
The Chimes: correct author Dickens ranked 2
The Circular Staircase: correct author Rinehart ranked 1
The Confession: correct author Rinehart ranked 1
The Cricket on the Hearth: correct author Dickens ranked 1
The Crimson Fairy Book: correct author Lang ranked 1
The Crown of Life: correct author Gissing ranked 1
The Cruise ofthe Snark: correct author London ranked 1
The Dawn of a Tomorrow: correct author Burnett ranked 17
The Day's Work: correct author Kipling ranked 1
The Deerslayer: correct author Cooper ranked 1
The Descent of Man: correct author Darwin ranked 1
The Diary of an Old Soul: correct author MacDonald ranked 3
The Disappearance of Lady Frances Carfax: correct author Doyle ranked 1
The Drums of Jeopardy: correct author MacGrath ranked 1
The Dutch Twins: correct author Perkins ranked 1
The Earth Trembled: correct author Roe ranked 1
The Emancipated: correct author Gissing ranked 1
The Errand Boy: correct author Alger ranked 1
The Eskimo Twins: correct author Perkins ranked 2
The Evil Genius: correct author Collins ranked 1
The Expression of the Emotions in Man and Animals: correct author Darwin ranked 5
The Fair Haven: correct author Butler ranked 1
The Faith of Men: correct author London ranked 1
The First Men in the Moon: correct author Wells ranked 1
The Flying U's Last Stand: correct author Bower ranked 1
The Fortunes and Misfortunes of the Famous Moll Flanders: correct author Defoe ranked 1
The Four Million: correct author Henry ranked 6
The Freedom of Life: correct author Call ranked 1
The French Revolution: correct author Carlyle ranked 3
The Garden of Allah: correct author Hichens ranked 8
The Golden Road: correct author Montgomery ranked 1
The Great Boer War: correct author Doyle ranked 3
The Great Stone Face: correct author Hawthorne ranked 1
The Guns of Bull Run: correct author Altsheler ranked 1
The Guns of Shiloh: correct author Altsheler ranked 1
The Haunted House: correct author Dickens ranked 19
The Heart of Rachael: correct author Norris ranked 1
The Heritage of the Souix: correct author Bower ranked 1

The Hound of the Baskervilles: correct author Doyle ranked 1
The House of Pride and Other Tales of Hawaii: correct author London ranked 1
The Hunting Tower: correct author Buchan ranked 1
The Innocence of Father Brown: correct author Chesterton ranked 1
The Iron Heel: correct author London ranked 1
The Island of Doctor Moreau: correct author Wells ranked 1
The Japanese Twins: correct author Perkins ranked 1
The Journal to Stella: correct author Swift ranked 3
The Jungle Book: correct author Kipling ranked 2
The Lady of the Lake: correct author Scott ranked 1
The Last of the Mohicans: correct author Cooper ranked 1
The Legacy of Cain: correct author Collins ranked 1
The Legend of Sleepy Hollow: correct author Irving ranked 1
The Lost Prince: correct author Burnett ranked 21
The Lure of the Dim Trails: correct author Bower ranked 1
The Man Who Knew Too Much: correct author Chesterton ranked 2
The Man Who Was Thursday: correct author Chesterton ranked 1
The Man Who Would Be King: correct author Kipling ranked 29
The Marble Faun Volume 1: correct author Hawthorne ranked 2
The Marble Faun Volume 2: correct author Hawthorne ranked 1
The Mirror of the Sea: correct author Conrad ranked 1
The Mutiny of the Elsinore: correct author London ranked 1
The Mystery of a Hansom Cab: correct author Hume ranked 1
The Nether World: correct author Gissing ranked 1
The New Machiavelli: correct author Wells ranked 1
The Night Born: correct author London ranked 1
The Odd Women: correct author Gissing ranked 1
The Orange Fairy Book: correct author Lang ranked 1
The Parent's Assistant: correct author Edgeworth ranked 1
The Path of the King: correct author Buchan ranked 1
The People of the Abyss: correct author London ranked 7
The Pickwick Papers: correct author Dickens ranked 1
The Pioneers: correct author Cooper ranked 2
The Princess and Curdie: correct author MacDonald ranked 2
The Princess and the Goblin: correct author MacDonald ranked 4
The Puppet Crown: correct author MacGrath ranked 1
The Red Fairy Book: correct author Lang ranked 1
The Red Planet: correct author Locke ranked 1
The Research Magnificent: correct author Wells ranked 2
The Return of Sherlock Holmes: correct author Doyle ranked 1
The Rich Mrs. Brgoyne: correct author Norris ranked 1
The Rose in the Ring: correct author McCutcheon ranked 1
The Ruling Passion: correct author Van Dyke ranked 1
The Scouts of Stonewall: correct author Altsheler ranked 1
The Scouts of the Vally: correct author Altsheler ranked 1
The Sea Wolf: correct author London ranked 1
The Second Jungle Book: correct author Kipling ranked 5
The Secret Garden: correct author Burnett ranked 32
The Secret of the Night: correct author Leroux ranked 1
The Secret Passage: correct author Hume ranked 1
The Shuttle: correct author Burnett ranked 9
The Sketch-Book of Geoffrey Crayon: correct author Irving ranked 1
The Son of the Wolf: correct author London ranked 3
The Spell of Egypt: correct author Hichens ranked 15
The Story of Julia Page: correct author Norris ranked 1

The Street of Seven Stars: correct author Rinehart ranked 1
The Swiss Twins: correct author Perkins ranked 1
The Talisman: correct author Scott ranked 1
The Thirty-Nine Steps: correct author Buchan ranked 1
The Time Machine: correct author Wells ranked 1
The Trail of the White Mule: correct author Bower ranked 1
The Two Destinies: correct author Collins ranked 1
The Uncommercial Traveller: correct author Dickens ranked 1
The Violet Fairy Book: correct author Lang ranked 1
The Vital Message: correct author Doyle ranked 1
The White People: correct author Burnett ranked 21
The Wisdom of Father Brown: correct author Chesterton ranked 1
The Woman in White: correct author Collins ranked 1
The World Set Free: correct author Wells ranked 1
The Yellow Fairy Book: correct author Lang ranked 1
This Side of Paradise: correct author Fitzgerald ranked 1
Those Extraordinary Twins: correct author Twain ranked 1
Through the Looking Glass: correct author Carroll ranked 1
Timothy Crump's Ward: correct author Alger ranked 1
Tish: correct author Rinehart ranked 1
Tom Swift Among the Diamond Makers: correct author Appleton ranked 1
Tom Swift and his Air Glider: correct author Appleton ranked 1
Tom Swift and his Air Scout: correct author Appleton ranked 1
Tom Swift and his Undersea Search: correct author Appleton ranked 1
Tom Swift in the Caves of Ice: correct author Appleton ranked 1
Tono Bungay: correct author Wells ranked 1
Tow Swift Among the Fire Fighters: correct author Appleton ranked 1
Twelve Stories and a Dream: correct author Wells ranked 1
Twixt Land and Sea: correct author Conrad ranked 1
Typee: correct author Melville ranked 1
Under the Lilacs: correct author Alcott ranked 1
Undertow: correct author Norris ranked 1
Viola Gwyn: correct author McCutcheon ranked 1
Voice of the City: correct author Henry ranked 4
Volcanic Islands: correct author Darwin ranked 1
Waifs and Strays Part 1: correct author Henry ranked 7
Walking: correct author Thoreau ranked 1
Waverly: correct author Scott ranked 1
What Can She Do: correct author Roe ranked 1
What's Bred in the Bone: correct author Allen ranked 3
When a Man Marries: correct author Rinehart ranked 1
When the Sleeper Wakes: correct author Wells ranked 1
Where Angels Fear to Tread: correct author Forster ranked 1
Where there's a Will: correct author Rinehart ranked 1
White Lies: correct author Reade ranked 1
Wieland: correct author Brown ranked 1
Within the Tides: correct author Conrad ranked 1
Wives and Daughters: correct author Gaskell ranked 1
Women in Love: correct author Lawrence ranked 40
Work- A Story of Experience: correct author Alcott ranked 1

BIBLIOGRAPHY

Bach, J.; Witten, I. "Lexical Attraction for Text Compression." *Data Compression Conference*. Snowbird, Utah, 29-31 March, 1999.

Bosch, R. A., Smith, J. A. "Separating Hyperplanes and the Authorship of the Disputed Federalist Papers," *American Mathematical Monthly*, v. 105, num. 7, pp. 601-608, 1998.

Diedrich, J., Kindermann, J., Leopold, E., and Paass, G. "Authorship Attribution with Support Vector Machines." Submitted to *Applied Intelligence*, 2000.

Efron, B.;Thisted, R. "Estimating the number of unseen species: How many words did Shakespeare know?" *Biometrika*, v. 63, pp. 435- 448, 1976.

Ephratt, M. "Authorship attribution - the case of lexical innovations." *Joint International Conference of the Association for Computers and the Humanities and the Association for Literary & Linguistic Computing*, 1997.

Farringdon, Jill M. *Analysing for Authorship: A Guide to the Cusum Technique*. Cardiff: University of Wales Press, 1996.

Foster, D. "Author Unknown: On the Trail of Anonymous," Henry Holt, New York, 2000.

Holmes, D.I.. Forsyth, R.S. "The Federalist Revisited: New Directions in Authorship Attribution". *Literary and Linguistic Computing*, v. 10, num. 2, pp. 111-27, 1995.

Holmes, D. I. "Stylometry: Its Origins, Development and Aspirations," presented to the *Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing,* Queen's University, Kingston, Ontario, 1997.

Holmes, D.I. "The Evolution of Stylometry in Humanities Scholarship." *Literary and Linguistic Computing*, v. 13, num. 3, pp. 111-117, 1998.

Karlgren, J., and Cutting, D. "Recognizing Text Genres with Simple Metrics using Discriminant Analysis." *Proceedings of the 15th. International Conference on Computational Linguistics*,: correct author COLING], Kyoto, 1994.

Lewis, D.D. "Feature Selection and Feature Extraction for Text Categorization." *Procceedings of the Speech and Natural language Workshop*, pp 212-217, February 1992.

Martindale, C., and McKenzie, D. "On the Utility of Content Analysis in Authorship Attribution: *The Federalist*." *Computers and the Humanities*, v. 29, pp 259-270, 1995.

Mosteller, F., and Wallace, D. L. *Inference and Disputed Authorship: The Federalist*. Springer-Verlag, New York, 1964.

Rudman, J. "The State of Authorship Attribution Studies: Some Problems and Solutions." *Computers and the Humanities*, v. 31, pp 351-365, 1998.

Shannon, C. E.. A Mathematical Theory of Communication. *The Bell System Technical Journal*, v. 27*,* 1948.

Tweedie, F.J.; Singh, S.; Holmes, D.I. "Neural Network Applications in Stylometry: The Federalist Paper." *Computers and the Humanities*, v. 30, pp. 1-10, 1996

Yuret, D. "Discovery of linguistic relations using lexical attraction." PhD Thesis, Department of Computer Science and Electrical Engineering, MIT. May, 1998.