Literary and Linguistic Computing Advance Access published July 26, 2007

Quantitative Authorship Attribution: An Evaluation of Techniques

Jack Grieve
English Department
Northern Arizona University

Abstract

The basic assumption of quantitative authorship attribution is that the author of a text can be selected from a set of possible authors by comparing the values of textual measurements in that text to their corresponding values in each possible author's writing sample. Over the past three centuries, many types of textual measurements have been proposed, but never before have the majority of these measurements been tested on the same dataset. A large-scale comparison of textual measurements is crucial if current techniques are to be used effectively and if new and more powerful techniques are to be developed. This article presents the results of a comparison of thirty-nine different types of textual measurements commonly used in attribution studies, in order to determine which are the best indicators of authorship. Based on the results of these tests, a more accurate approach to quantitative authorship attribution is proposed, which involves the analysis of many different textual measurements.

Correspondence:
Jack Grieve, 520 South Leroux, Flagstaff, AZ 86001, USA
E-mail: jwg39@nau.edu

1 Introduction

Since the late 19th century, quantitative methods have been used to determine the author of anonymous texts. In quantitative authorship attribution, the values of textual measurements in the anonymous text are compared to their corresponding values in a series of possible author writing samples, in order to determine which possible author writing sample is the best match. While investigators of authorship have proposed many textual measurements over the past three centuries, never before has a large-scale evaluation of these measurements been conducted. Such a comparison is long overdue: if we are to resolve current cases of disputed authorship and develop new and more powerful techniques, then we must know which of our measurements are most useful for attributing authorship. The goal of this study is thus to evaluate thirty-nine commonly used types of textual measurements, in order to determine which are the best indicators of authorship. In addition, based on the results of these tests, a general approach to quantitative authorship attribution is proposed, which involves the simultaneous analysis of many different types of textual measurements.

2 Attribution Algorithms

To conduct a fair comparison of the most commonly used sets of textual measurements in quantitative authorship attribution it is necessary that each set of measurements be inserted into an otherwise identical attribution algorithm and tested on the same dataset. In this section, the basic quantitative attribution algorithm is described, as well as the thirty-nine sets of textual measurements that will take their turn at its core. The evaluation procedure is described in Section 3.

doi:10.1093/llc/fqm020

2.1 Input

An attribution algorithm compares an anonymous text to a series of possible author writing samples. As such, it takes as input an anonymous text and a set of possible author writing samples. Generally, each possible author is represented by multiple writing samples so as to increase the likelihood that any patterns found in the sample are characteristic of that author.

2.2 Textual measurements

In order to determine which possible author's writing sample most resembles the anonymous text, the attribution algorithm compares the values of a set of textual measurements in the anonymous text to their corresponding values in the possible author writing samples. The thirty-nine sets of textual measurements tested here are defined below. In these definitions, a *character* is an indivisible textual unit, including graphemes, digits, punctuation marks, and whitespaces; a *grapheme* is a letter of the alphabet; a *word* is a continuous string of graphemes and/or digits; a *sentence* is a continuous string of characters, excluding question marks, exclamation marks, newlines and nonabbreviatory periods; a *n-word collocation* is a sequence of *n* words; an *n-gram* is a sequence of *n* characters; and a *profile* is a set of textual measurements.

2.2.1 *Word-length*

Two word-length measurements are tested. The first measurement is average word-length, which is calculated by dividing the total number of digits and graphemes in a text by the total number of words. The second measurement is a word-length distribution, which consists of the relative frequency of 1-character words, 2-character words, etc. in a text, where the relative frequency of each word-length is calculated by dividing the total number of words of that length in the text by the total number of words. Various forms of this measurement are tested, which differ in terms of the range of word-lengths that the distribution spans.

Attribution studies that consider word-length include Brinegar (1963), Foster (1989), Forsyth *et al.* (1999), Fucks (1952, 1954) Fucks and Lauter (1965), Mendenhall (1887, 1901), O'Donnell (1966), Radday (1970), Smith (1983), and Williams (1970).

2.2.2 *Sentence-length*

Four sentence-length measurements are tested. The first measurement is average sentence-length in words, which is calculated by dividing the total number of words in a text by the total number of sentences. The second measurement is a sentence-length distribution in words, which consists of the relative frequency of 1-word sentences, 2-word sentences, etc. in a text, where the relative frequency of each sentence-length is calculated by dividing the total number of sentences of that length in the text by the total number of sentences. Various forms of this measurement are tested, which differ in terms of the range of sentence-lengths that the distribution spans. The third measurement is average sentence-length in characters, which is calculated by dividing the total number of characters in a text by the total number of sentences. The fourth measurement is a sentence-length distribution in characters, which consists of the relative frequency of, for example, 1-to-10-character sentences, 11-to-20-character sentences, etc. in a text, where the relative frequency of a sentence-length-range is calculated by dividing the total number of sentences of those lengths in the text by the total number of sentences. Various forms of this measurement are tested, which differ in terms of the range of sentence-lengths that the distribution spans.

Attribution studies that consider sentence-length include Eddy (1887), Herdan (1960, 1965), Kjetsaa (1978), Mascol (1888a, 1888b), Morton (1965), Radday (1970), Wake (1957), Williams (1940), and Yule (1939).

2.2.3 *Vocabulary richness*

Eleven vocabulary richness measurements are tested. The formulae for ten of these measurements are presented below, where N is the total number of words in a text (i.e. word tokens), V is total number of vocabulary items in a text (i.e. word types), V_i is the total number of vocabulary items that occur exactly i-times in a text, p_v is the relative frequency of the v-th most frequent

vocabulary item in a text, and a is an arbitrary constant.

$$\text{Type}-\text{Token} = V/N \quad (1)$$

$$K = 10^4(\Sigma i^2 V_i - N)/N^2 \quad (2)$$

$$R = V/\sqrt{N} \quad (3)$$

$$C = \log V/\log N \quad (4)$$

$$H = (100 \log N)/(1 - V_1/V) \quad (5)$$

$$S = V_2/V \quad (6)$$

$$k = \log V/\log(\log N) \quad (7)$$

$$LN = (1 - V^2)/(V^2 \log N) \quad (8)$$

$$\text{Entropy} = -100\Sigma p_v \log p_v \quad (9)$$

$$\text{W} = N^{V-a} \quad (10)$$

The eleventh vocabulary richness measurement is a limited Type–Token Ratio, which is based on only the first n-number of words in every text, where n is the number of words in the shortest writing sample. This measurement is made because the Type–Token Ratio is known to be very sensitive to text-length—as a text gets longer, new word-types are introduced at a slower rate.

Attribution studies that consider vocabulary richness include Baayen *et al.* (1996), Chaski (2001), Holmes (1992), Holmes and Forsyth (1995), Kjetsaa (1978), Pollatschek and Radday (1981, 1985), Radday (1970), Somers and Tweedie (2003) Tweedie and Baayen (1998), and Yule (1944).

2.2.4 *Grapheme frequency*

Four grapheme frequency measurements are tested. The first measurement is a simple grapheme profile, which consists of the relative frequency of the twenty-six graphemes of the English alphabet, where the relative frequency of each grapheme is calculated by dividing frequency of that grapheme in a text by the total number of graphemes. The second measurement is a single-position grapheme profile, which consists of the relative frequency of graphemes occurring in a particular position in the words of a text (e.g. first grapheme, second grapheme, last grapheme in a word), where the relative frequency of each grapheme occurring in that particular position is calculated by dividing the frequency of that grapheme in that position in the text by the total number of words that contain that position. The third measurement is a multiposition grapheme profile, which consists of multiple single-position grapheme profiles (e.g. first three positions of a word). Various forms of both word-position measurement are tested, which differ in terms of which word-positions are being analyzed. The fourth measurement is a word-internal grapheme profile, which consists of the percentage of words in a text that contain each of the twenty-six graphemes of the English language, where the percentage of words in the text that contain a particular grapheme is calculated by dividing the number of words in the text that contain at least one instance of that grapheme by the total number of words.

Attribution studies that consider grapheme frequency include Herdan (1966), Ledger (1995), Ledger and Merriam (1994), Merriam (1988, 1994, 1998), and Yule (1944).

2.2.5 *Word frequency*

Three word frequency measurements are tested. The first measurement is a simple word profile, which consists of the relative frequency of a set of high frequency words, where the relative frequency of each word is calculated by dividing the frequency of that word in a text, by the total number of words. Various forms of the simple word profile are tested here, which differ in terms of the minimum frequency cut off for a word to be included in the profile. The second measurement is a single-position word profile, which consists of the relative frequency of a set of words occurring in a particular position in the sentences of a text (e.g. first word, second word, and last word of a sentence), where the relative frequency of each word occurring in that particular position is calculated by dividing the frequency of that word in that position in the text by the total number of sentences that contain that position. The third measurement is a multi-position word profile, which consists of multiple single-position word profiles (e.g. first four words in a sentence). Various forms of both sentence-position measurements are tested, which differ in terms of which sentence-positions are being analyzed.

Attribution studies that consider word frequency include Burrows (1992), Burrows and Craig (2001), Burrows and Hassall (1988), Ellegård (1962a, 1962b), Holmes *et al.* (2001), Kenny (1978), Levison *et al.* (1966), Mascol (1888a, 1888b), Michaelson and Morton (1972), Morton (1965, 1978), Morton and Levison (1966), Morton and McLeman (1964, 1966), Mosteller and Wallace (1963, 1964, 1984), Smith (1991, 1992, 1993), and Tweedie *et al.* (1998).

2.2.6 *Punctuation mark frequency*

Five punctuation frequency measurements are tested. The first three measurements are variants of a simple punctuation mark profile. These measurements consists of the relative frequency of eight punctuation marks (. , : ; - ? (‘), but differ in how these relative frequencies are calculated. In the first measurement, the relative frequency of each punctuation mark is calculated by dividing the frequency of that punctuation mark in a text by the total number of characters; in the second measurement, by the total number of punctuation marks; and, in the third measurement, by the total number of words. The fourth measurement is a punctuation and grapheme profile. The fifth measurement is a punctuation and word profile. These measurements are combinations of the punctuation profile and the grapheme and word profiles.

Attribution studies that consider punctuation mark frequency include Chaski (2001), Mascol (1888a, 1888b), and O’Donnell (1966).

2.2.7 *Collocation frequency*

Two collocation measurements are tested. The first measurement is a 2-word collocation profile, which consists of the relative frequency of a set of high frequency two-word collocations, where the relative frequency of a collocation is calculated by dividing the frequency of that collocation in a text, by the total number of two-word collocations. The second measurement is a 3-word collocation profile.

Attribution studies that consider collocation frequency include Hoover (2002), Merriam (1979, 1980, 1982), Morton (1978), and O’Brien and Darnell (1982).

2.2.8 *Character-level n-gram frequency*

Eight character-level *n*-gram measurements are tested. The first measurement is a 2-gram profile, which consists of the relative frequency of a set of high frequency 2-grams, where the relative frequency of a 2-gram is calculated by the dividing the frequency of that 2-gram in a text, by the total number of 2-grams. The other seven measurements are 3- through 9-gram profiles. Various forms of the *n*-gram profiles are tested, which differ in terms of the minimum frequency cut off for an *n*-gram to be included in the profile.

Attribution studies that consider *n*-gram frequency include Bennett (1976), Clement and Sharp (2003), Keselj *et al.* (2003), and Peng *et al.* (2003).

2.3 Comparison and output

As outlined earlier, the basic attribution algorithm takes as input an anonymous text and a set of possible author writing samples. The algorithm then reduces each input text to a set of textual measurements plus their specific values in that text. Because each author is represented by multiple writing samples, these sets of measurements are then combined to form one set of measurements for each possible author. This is accomplished by averaging the values of each textual measurement across each of that author’s writing samples. Finally, the algorithm compares the values of the measurements in the anonymous text to their corresponding values for each possible author in order to determine which pair is the closest match.

The most common statistic used in authorship attribution to compare the values of a set of textual measurements is the chi-squared statistic. Chi-square is simple non-parametric goodness-of-fit statistic that is used to determine if a sample, represented by a set of observed frequencies (O), could have been drawn from a particular population, represented by a corresponding set of expected frequencies (E).

$$(11)\quad \chi^2 = \sum((O_i - E_i)^2/E_i) \quad i = 1, 2, 3, \ldots, n$$

The lower the chi-square value, the more confident one may be that the sample was drawn from that population: if the two sets are identical, then the

chi-square value is zero. To interpret a nonzero chi-square value, a critical chi-square table is consulted to determine if the value is low enough to reasonably conclude that that sample was drawn from that population.

In attribution studies, the chi-squared test is used to compare the observed frequencies of a set of textual measurements in an anonymous text to the sets of frequencies that would be expected if the text were written by a particular possible author, based on an analysis of that author's writing samples (e.g. Brinegar 1963; Chaski 2001; Forsyth and Holmes 1996; Kenny 1978; Morton 1965; O'Brien and Darnell 1982; Usher and Najock 1982). In this study, the chi-squared test is used specifically to compare the observed values of the set of measurements in the anonymous text to their corresponding values in each possible author writing sample. The algorithm then outputs a list of possible authors ranked by ascending chi-square value, where the author associated with the smallest chi-square value is deemed to be the anonymous text's best match. The critical chi-square table is not consulted: the algorithm outputs a ranking of possible authors, which is interpreted as an ordered list of most likely authors.

3 Evaluation

To evaluate the thirty-nine sets of textual measurements defined earlier, it is necessary that each set be inserted into an identical attribution algorithm and tested in an identical manner on the same dataset. In this section, the evaluation procedure used in this study is described: the corpus of possible authors is presented and the way in which this dataset was used to test the performance of the attribution algorithms is explained.

3.1 Corpus compilation

A corpus of possible authors consists of a collection of author-based corpora, each of which represents a variety of language in which a particular possible author writes. This corpus will provide both the writing samples and the 'anonymous' test texts upon which the attribution algorithms will be tested. There are two main considerations when compiling a corpus of possible authors. First, highly representative author-based corpora must be compiled. Second, when these author-based corpora are combined, the corpus of possible authors must also constitute a highly representative corpus in and of itself.

3.1.1 *Author-based corpus compilation*

It is not a trivial matter to define the variety of language in which an author writes. Most writers interact with multiple readers, at multiple times, and in multiple registers, and so one must decide which of an author's many varieties the author-based corpus will represent. When attributing an anonymous text, it is unnecessary and unsound to compile an author-based corpus that attempts to represent the variety that encompasses all that author's written utterances: the anonymous text is the product of a single situation and so each author-based corpus should be composed of texts produced in the most similar register, for the most similar audience, and around the same point in time as the anonymous text. Otherwise, the investigator might get false negatives: when the anonymous text is compared to the writings of its author they may not match because of variation that is the product of differences in audience or register or time. Rather, it is best to compile author-based corpora that represent the narrowest variety of language possible. This is accomplished here by controlling the dialect, register and era of the author-based corpora.

Clearly, the dialect of each of the author-based corpora must be defined in terms of a single author, but to define the narrowest of dialects it is also necessary to specify a stable audience. The indivisible dialect produced by a single speaker for a single audience is known as an *idiolect*. This term was introduced by Bernard Bloch (1948), and while it is often used by linguists (e.g. Hockett, 1958) to refer to the variety of language that encompasses the totality of an individual's utterances, this is not how Bloch intended the term to be used. Rather, he defined an idiolect as "the totality of the possible

utterances of one speaker at one time in using a language to interact with one other speaker'' (1948:7). In this study, each author-based corpus represents an idiolect. This was accomplished by only selecting authors who write for the London *Telegraph*, and by only sampling these authors' regular *Telegraph* opinion columns. Admittedly, the readership of a newspaper column is never entirely stable, but because the readership is fairly stable and because the readership is so large and anonymous that the columnist never knows its exact composition, it is assumed that the columnist will usually treat his readership as a stable audience, especially in cases of established columnists who write for major newspapers with a large and dedicated audience, such as the *Telegraph*.

By choosing to compile author-based corpora that represent the variety of language in which *Telegraph* opinion columnists write, the register of these corpora has also been defined in very narrow terms: the *Telegraph* opinion newspaper column is a very specific type of register. This register is particularly well-suited for attribution studies because newspaper columns are plentiful and in the public domain, and because the *Telegraph* offers a large online archive from which texts can be freely downloaded. The newspaper opinion column register also provides texts that are carefully written, of a comparable length, and that are short enough (usually 500–2,000 words) to provide a challenging test for the algorithms.

It is also necessary to minimize the time span from which the texts are gathered in order to limit temporal variation. Fortunately, newspaper columnists are also some of the most prolific writers of published English texts, and so when compiling this corpus it was possible to include a relatively a large number of texts in each author-based corpus, while keeping the time span relatively short. In particular, each author-based corpus contains forty columns. For most columnists this requires that texts be sampled from a one-year time span. Usually this time span ranges from January 2004 to January 2005, but in all cases the columns were written between the years 2000 and 2005.

3.1.2 *Corpus of possible authors compilation*

To test the performance of an attribution algorithm it is not enough to assemble a set of highly representative author-based corpora: the set of author-based corpora itself must be highly representative of some variety of language. The more representative this corpus of possible authors is, the more realistic and challenging a test it will provide for the attribution algorithms.

In an actual case of disputed authorship, the dialect, era and register of the anonymous text should direct the compilation of the author-based corpora. For example, if an investigator is attributing an eighteenth century Scottish poem, then the investigator should only consider poems written by eighteenth century Scottish poets. When the author-based corpora are combined, the resultant corpus of possible authors will thus naturally be highly representative of some variety of language, in this case, eighteenth century Scottish poetry. However, in a study such as this, there is no anonymous text to direct the compilation of the author-based corpora. Nonetheless, when these author-based corpora are combined they must still constitute a highly representative corpus of possible authors, otherwise an algorithm that appears to be identifying authorship may instead be identifying some other feature of the extra-linguistic situation in which the text was produced. For example, if each of the authors originated from a different region or wrote in a different register or at a different time, then it would be impossible to know if an algorithm that tested successfully was identifying authorship or dialect or register or era. Similarly, the corpus of possible authors must be controlled for subject: if the corpus of possible authors is to provide a valid test, then each possible author must write about a similar range of topics. Otherwise, an algorithm capable of topic-based text classification would appear to be capable of authorship-based text classification.

In this study, a highly representative corpus of possible authors was compiled by combining forty author-based corpora that contain texts written by authors from similar social backgrounds (middle-aged, conservative, Anglo-Saxon, upper-middle-class, well-educated, British), which are

written in the same register (*Telegraph* opinion column) and for the same audience (the readership of the *Telegraph*'s opinion section), and which are published over the same short span of time (2000–2005). The likelihood that the topics of the texts cluster by authorship has been minimized because newspaper opinion columnists, especially when writing at the same time and in the same city, will tend to write about a similar range of subjects.

Overall, the least successful control has been the social backgrounds of the authors, for while most of the columnists are middle-aged well-educated Britons, some of the columnists—such as Barbara Amiel, who is a Canadian, Zoe Heller, who lives in New York, and W. F. Deeds, who is in his nineties—are from different social backgrounds. Stricter controls, however, would have made it impossible to include forty possible authors in the corpus. It is important the corpus contains such a large number of authors for two reasons: it allows for the attribution algorithms to be tested on a large number of possible authors simultaneously—most attribution algorithms have never been asked to distinguish between forty possible at once—and it allows for attribution algorithms to be tested on multiple smaller sets of possible authors, thereby increasing the accuracy of the tests.

Table 1 presents the corpus of possible authors used in this study. For each author-based corpus, the table lists its author's name, the time span over which its texts were written, the total number of words, and the basic subjects that its texts discuss (B: *British Politics*, W: *World Affairs*, C: *Culture*, A: *Art*, S: *Sport*, E: *Economics*, R: *Religion*, H: *Health*). In total, the *Telegraph Columnist Corpus* contains 1.5 million words spread out over 1600 texts written by forty authors.

3.2 Algorithm evaluation

The corpus of possible authors described earlier is used to test the performance of the attribution algorithms in this study as follows. First, one author is selected from the set of possible authors. Second, one test text (i.e. acting as the anonymous text) is selected from that author's corpus. Third, the attribution algorithm attributes the test text by comparing the test text, as described earlier, to the remaining texts in each possible author's corpus (i.e. acting as the writing samples). The test text is then returned to its author-based corpus and the procedure is repeated with a new test text. Once all the texts in that author based corpus have been attributed, the procedure is repeated with the next author-based corpus. Once all the texts in all the author-based corpora in the corpus of possible authors have been attributed, the attribution algorithm's success rate is calculated by dividing the number of successful attributions by the total number of attributions attempted.

In this study, each attribution algorithm is subjected to seven tests. These tests all conform to the basic testing procedure described earlier, but vary in terms of the number (forty, twenty, ten, five, four, three, and two) of possible authors included in the corpus of possible authors. Except for the test involving the full set of forty possible authors, not all the possible authors will be used in any one running of the tests, and therefore the attribution algorithms can be tested on multiple sets of possible authors, so that more accurate results may be obtained. Specifically, all the tests conducted in this study that involve fewer than forty possible authors were repeated 200 times, using 200 different sets of possible authors drawn randomly from the complete set of forty possible authors. This particular number of permutations was chosen by subjecting various algorithms to tests differing only in the value of this one parameter: it was found that 200 permutation tests yielded results within 0.5% of the results of 1,000 and 2,000 permutation tests. When an algorithm is tested over multiple permutations of possible authors, its overall success rate is calculated by averaging its success rates over each permutation. In order to ensure that the results of these tests are commensurable, the same 200 randomly generated sets of possible authors are used every time an algorithm is subjected to these tests.

4 Results

All of the results presented here take the form of attribution algorithm accuracy tables. Each table

Table 1 The *Telegraph* Columnist Corpus

Name	Date	Word	Subject
Barbara Amiel	May 2003–May 04	47,715	WPCR
Craig Brown	Jul 2004–Jan 05	37,860	CPAW
Alexander Chancellor	Apr 2003–Nov 04	40,844	CPAWR
Ross Clark	Feb 2004–Jan 05	31,197	EPWCS
Neil Collins	May 2003–Nov 04	40,318	PEC
Janet Daley	Jan 2004–Dec 04	39,380	PWCE
Theodore Dalrymple	Apr 2001–Jan 05	38,504	HCPW
Matthew d'Ancona	Mar 2004–Jan 05	53,891	PW
W.F. Deeds	Jan 2004–Dec 04	26,300	PWCS
Nigel Farndale	Jan 2004–Jan 05	33,024	PCWAS
Zoe Heller	May 2003–Jul 04	41,893	CPW
Susannah Herbert	Oct 2002–Jan 05	36,900	PWC
Christopher Howse	Jun 2004–Jan 05	30,218	RCAPW
Armando Iannucci	Dec 2002–Jul 04	31,445	PWC
Boris Johnson	Mar 2004–Jan 05	41,221	PCWE
Daniel Johnson	Oct 2001–Nov 04	41,078	CPWARE
Frank Johnson	Nov 2003–Jan 05	24,750	PCW
John Keegan	May 2002–Jan 05	44,028	WPC
Sam Leith	May 2004–Dec 04	26,605	CPASW
Jemima Lewis	Mar 2004–Jan 05	36,311	CWPA
Andrew Marr	Dec 2003–Dec 04	26,798	PWEC
Jenny McCartney	Mar 2004–Jan 05	37,472	CPAW
Charles Moore	Sep 2003–Jan 05	48,684	PWC
Harry Mount	May 2002–Jan 05	31,472	CAPW
Kevin Myers	Mar 2004–Jan 05	38,463	CAPWS
Adam Nicolson	Dec 2003–Jan 05	38,510	CWP
Alasdair Palmer	Jun 2002–Jan 05	37,737	WPC
Stephen Pollard	May 2001–Jan 05	35,538	WCP
Oliver Pritchett	Jul 2004–Dec 04	31,034	CPW
Anne Robinson	Apr 2003–May 04	40,982	CAPW
Stephen Robinson	Dec 2003–Dec 04	36,014	CWPS
Sarah Sands	Mar 2004–Jan 05	36,285	CPSWA
Peter Simple	Oct 2003–Jan 05	33,741	CPS
Mark Steyn	Apr 2004–Dec 04	45,499	CPWE
Rachel Sylvester	Sep 2003–Jan 05	38,929	PWE
Alice Thompson	Dec 2003–Jan 05	41,598	PWCE
George Trefgarne	Sep 2003–Jan 05	39,328	EPW
Tom Utley	Mar 2004–Jan 05	43,453	PCWAE
Jim White	Mar 2004–Dec 04	39,375	CAPS
Vicki Woods	Feb 2004–Dec 04	27,916	WPC

presents the results of subjecting multiple attribution algorithms (defined in terms of sets of textual measurements), to multiple tests (defined in terms of the number of possible authors per permutation). The result of subjecting a particular attribution algorithm to a particular test is recorded in the cell at the intersection of the test's row and the algorithm's column, as the percentage of texts correctly attributed. These results are interpreted in two ways. First, the relative accuracy of the various algorithms is considered. Second, if a particular attribution algorithm achieves at least 75% accuracy on a particular test, then that algorithm is deemed to have performed successfully on that test.

4.1 Word- and sentence-length

Table 2 presents the results of testing the attribution algorithms based on measurements of word- and sentence-length.

Table 2 Word- and sentence-length results

Textual measurement			Test accuracy (%)						
Type	Variant		Possible authors						
	Unit	Range	40	20	10	5	4	3	2
Average word-length	Grapheme		7	12	22	39	46	55	70
Average sentence-length	Word		6	11	21	37	44	53	69
Average sentence-length	Grapheme		6	12	22	39	45	53	70
Word-length profile	one grapheme	1–15 characters	18	26	39	54	60	68	79
Word-length profile	one grapheme	1–5 characters	11	18	29	45	51	60	74
Sentence-length profile	five words	1–50 words	11	18	29	44	51	60	74
Sentence-length profile	five words	1–30 words	8	16	26	41	47	57	71
Sentence-length profile	ten words	1–50 words	10	17	28	44	50	59	73
Sentence-length profile	ten words	1–30 words	8	14	24	38	45	54	70
Sentence-length profile	twenty-five characters	1–300 characters	12	20	31	46	53	62	74
Sentence-length profile	twenty-five characters	1–200 characters	10	17	28	43	50	59	73
Sentence-length profile	fifty characters	1–300 characters	11	19	30	45	52	61	74
Sentence-length profile	fifty characters	1–200 characters	9	16	26	41	48	57	72

The first three algorithms, which are based on the value of a single measurement of average word- or sentence-length, would appear to be of little use to investigators of authorship. The multivariate algorithms did not perform much better: only the larger variant of the word-length distribution algorithm proves to be capable of distinguishing between two possible authors with any degree of success. Of the two types of multivariate sentence-length algorithms tested here, those that measure sentence-length in characters were slightly more successful than those that measure sentence-length in words. This is unremarkable because only the character-based measurements are sensitive to word- as well as sentence-length. Overall, the larger multivariate algorithms were more successful than those based on fewer measurements. Finally, there may be many reasons why the word-length algorithms have achieved slightly better results than the sentence-length algorithms, but perhaps the most important is that a text is composed of far more words than sentences, and hence any measurement of word-length will be based on far more observations than any measurement of sentence-length.

4.2 Vocabulary richness

Table 3 presents the results of testing the attribution algorithms based on measurements of vocabulary richness.

The unrestricted Type–Token ratio has clearly outperformed the restricted Type–Token ratio (which is based on the first 119 words of each text), yet the unrestricted Type–Token ratio achieves acceptable results only when asked to distinguish between two possible authors. Yule's *K* and Simpson's *D* (which are functionally equivalent) are even less successful. The poor performance of these measurements is noteworthy because, along with the restricted Type–Token ratio, they are the only vocabulary richness measurements that are theoretically stable across texts of different lengths (Tweedie and Baayen, 1998). This unexpected result would seem to be a result of the fact that text-length is itself a mediocre indicator of authorship over this corpus of possible: when text-length was tested *post hoc* as an indicator of authorship, the algorithm achieved 77%, 65%, 56%, 50%, 33%, 20%, and 11%, when asked to distinguish between two, three, four, five, ten, twenty, and forty possible authors. These results are better than any of the individual vocabulary richness algorithms, and likely due to the fact that newspaper columnists tend to write articles of a relatively stable length

Of the remaining measurements, Sichel's *S* and Michéa's *M* (which are functionally equivalent) and Honoré's *H* perform relatively poorly, whereas entropy and *W* and the various logarithmic

Table 3 Vocabulary richness results

Textual measurement	Test accuracy (%)						
	Possible authors						
	40	20	10	5	4	3	2
Unrestricted Type–Token ratio	8	16	27	44	51	61	75
Restricted Type–Token ratio	3	7	14	27	33	42	59
Yule's *K* and Simpson's *D*	6	10	18	33	38	49	65
Guiraud's *R*	7	13	24	41	48	58	73
Herdan's *C*	7	14	25	42	49	59	73
Dugast's *k*	8	14	24	41	48	56	72
Honoré's *H*	7	13	23	38	45	54	70
Sichel's *S* and Michéa's *M*	4	9	16	29	35	45	61
Entropy	8	14	24	40	47	56	72
Tuldava's *LN*	11	18	31	49	55	64	77
W (a = −0.165)	11	17	26	40	46	53	68
W (a = −0.172)	11	17	26	40	45	52	67

Table 4 Grapheme frequency results

		Test accuracy (%)						
Textual measurement		Possible authors						
Type	Variant	40	20	10	5	4	3	2
Grapheme profile		25	35	47	62	67	74	83
Single-position grapheme profile	1st grapheme in word	20	30	41	56	62	69	80
Single-position grapheme profile	2nd grapheme in word	20	29	41	56	62	69	80
Single-position grapheme profile	3rd grapheme in word	16	24	35	49	55	63	75
Single-position grapheme profile	Last grapheme in word	27	36	49	63	68	73	84
Single-position grapheme profile	2nd to last graph in word	23	31	43	57	63	70	81
Single-position grapheme profile	3rd to last graph in word	19	28	41	56	61	69	80
Multiposition grapheme profile	1st three graphemes in word	34	44	56	69	73	79	87
Multiposition grapheme profile	1st six graphemes in word	43	53	64	76	79	84	90
Multiposition grapheme profile	Last three graphs in word	31	41	53	67	72	77	86
Multiposition grapheme profile	Last six graphs in word	42	52	63	74	79	83	90
Multiposition grapheme profile	First and last six graphs	49	58	68	79	82	86	92
Word-internal grapheme profile		28	39	51	65	70	76	85

attempts to stabilize the Type–Token ratio (Herdan's *C*, Guiraud's *R*, Dugast's *k*) all perform relatively well. In fact, the most successful of the vocabulary measurements is Tuldava's *LN*—a particularly complex logarithmic manipulation of the Type–Token ratio. Overall, measurements of vocabulary richness based on the entire (i.e. *K, D*) or part (*S, M, H*) of the grouped word frequency distribution are therefore less accurate than measurements based solely on the number of word-tokens and word-types in a text (*TTR, LN, C, R, k, W*). However, none of these algorithms are very successful, probably because all are based on the values of a single measurement, and because all are too sensitive to a text's subject matter for their value to remain stable across texts that range across many different topics.

4.3 Grapheme frequency

Table 4 presents the results of testing the attribution algorithms based on measurements of the relative frequency graphemes.

Table 5 Word frequency results

Textual measurement		Test accuracy (%) Possible authors						
Type	**Limit**	**40**	**20**	**10**	**5**	**4**	**3**	**2**
Word profile	In at least two texts per author	44	53	63	73	77	82	88
Word profile	In at least five texts per author	48	57	67	77	80	85	88
Word profile	In at least ten texts per author	45	54	64	75	79	84	90
Word profile	In at least fifteen texts per author	40	50	61	73	77	81	88
Word profile	In at least twenty texts per author	39	48	59	71	75	80	88
Word profile	In at least twenty-five texts per author	36	46	58	70	74	80	87
Word profile	In at least thirty texts per author	33	44	56	70	74	79	87
Word profile	In at least forty texts per author	16	23	35	50	57	64	57

The simple grapheme profile algorithm is more successful than any tested thus far, yet it still does not distinguish successfully between more than two possible authors. The various forms of the single-position grapheme algorithm are less accurate, except for the word-final grapheme algorithm, presumably because of its sensitivity to suffix usage. The multiposition grapheme algorithms are more successful: the 6- and 12-position algorithms achieve acceptable results when asked to distinguish between up to five possible authors. Once again, it would seem that the more measurements considered, the better the results. The word-internal grapheme profile also performs relatively well, successfully distinguishing between up to three possible authors.

The relative success of these algorithms is likely due to a combination of factors, such as an author's preference for particular sounds, spellings, synonyms, affixes, and words of particular etymological origins. But, perhaps the most important property of graphemes is that they are the most frequent potential indicator of authorship in any English text, and as such any patterns in their usage will have a better chance to emerge.

4.4 Word frequency

Table 5 presents the results of testing the attribution algorithms based on measurements of the relative frequency of words.

The eight algorithms tested here are based on variants of the basic word frequency profile. The largest word profile is the 2-limit profile, which consists of the 265 words that occur in at least two of every possible author's forty texts. All the other word frequency profiles are subsets of these 265 words: with each successive raising of the limit, a smaller set of words remains. The smallest word profile is the 40-limit profile, which contains only those five words (*and, the, to, a, of*) that occur in all 1,600 texts in the corpus. As the limit is raised and the profile shrinks, the content words are lost first. For example, the only content words left in the 10-limit profile are *made, said*, *time* and *people*, whereas roughly half of the original 265 words are content words.

The most accurate word frequency algorithms are based on the 5- and 10-limit word profiles. These two profiles contain most of the function words, but most of the content words have been stripped away. Both variants successfully distinguish between up to five possible authors, but the 10-limit variant performs slightly better on sets of two authors, whereas the 5-limit variant performs slightly better on all larger sets of possible authors. Higher limit algorithms do not fare as well, probably because function words, as opposed to content words, are being removed from the profiles. A common assumption of authorship attribution thus appears to be true: function words are better indicators of authorship than content words. For this reason, unlike most of the attribution algorithms tested in this study, larger word profiles do not lead to better results: the largest two-limit

Table 6 Punctuation mark frequency results

Textual measurement		Test accuracy (%) Possible authors						
Type	**Variant/limit**	**40**	**20**	**10**	**5**	**4**	**3**	**2**
Punctuation mark profile	By punctuation marks	30	40	53	67	71	77	86
Punctuation mark profile	By words	34	45	57	71	75	80	88
Punctuation mark profile	By characters	34	46	58	72	76	80	89
Grapheme and punctuation profile		50	60	70	81	84	87	93
Word and punctuation profile	In at least five texts per author	63	72	80	87	89	92	95
Word and punctuation profile	In at least ten texts per author	61	69	77	86	88	91	95
Word and punctuation profile	In at least twenty texts per author	57	66	75	80	83	87	94

variant does not perform as well as the smaller five-limit and ten-limit variants, presumably because of a higher percentage of content words.

4.5 Punctuation mark frequency

Table 6 presents the results of testing the attribution algorithms based on measurements of punctuation mark frequency.

When the relative frequency of each punctuation mark is calculated relative to the number of punctuation marks in the text, the method does not fare as well as when calculated relative to the total number of words or characters in the text. These attribution algorithms successfully distinguish between up to four possible authors. The combination algorithms are even more successful. The punctuation and grapheme algorithm distinguishes successfully between up to five possible authors and achieves 92% accuracy when distinguishing between sets of two possible authors. The punctuation and (5-limit) word algorithm performs even better: it is the most accurate individual algorithm tested in this entire study, successfully distinguishing between up to ten possible authors and achieving 95% accuracy when distinguishing between two possible authors.

These results are particularly impressive because there are only eight punctuation marks included in the punctuation profile, as opposed to, for example, the 264 graphemes-position pairs included in the largest multiposition grapheme profile. Overall, the frequency of individual punctuation marks is therefore one of the most potent quantitative indicators of authorship, despite the fact that this measurement has rarely been analyzed in attribution studies. Punctuation mark frequency is probably a good indicator of authorship because there is so much opportunity for variation in usage: an author can reasonably avoid using every punctuation mark save the period and perhaps the comma and question mark.

4.6 Positional stylometry

Table 7 presents the results of testing the attribution algorithms based on measurements of the relative frequency of words in particular sentence-positions, and the relative frequency of collocations. These methods were introduced by Andrew Morton and together are often referred to as *positional stylometry*.

The first algorithm tested here is based on the frequency of words that occur at the beginning of a sentence. This is one of the most successful word position algorithms, and yet it only achieves 75% accuracy when distinguishing between two possible authors. The other single-position algorithms are even less successful, most especially when words are counted in relationship to the end of the sentence. The multiposition profiles do not fare much better. The best of these algorithms is based on the frequency of words occurring in the first four positions of a text's sentences. Like measures of sentence-length, this lack of success is probably because these measurements are based on the frequency of fairly infrequent strings of characters.

Table 7 Positional stylometry results

Textual measurement		Test accuracy (%) Possible authors						
Type	**Variant**	**40**	**20**	**10**	**5**	**4**	**3**	**2**
Single-position word profile	1st word in sentence	17	30	36	50	56	64	75
Single-position word profile	2nd word in sentence	11	18	27	41	47	56	69
Single-position word profile	3rd word in sentence	7	13	21	35	41	50	64
Single-position word profile	4th word in sentence	6	10	17	30	35	45	59
Single-position word profile	Last word in sentence	4	7	13	25	30	39	56
Single-position word profile	2nd to last word in sentence	6	11	18	31	37	46	61
Single-position word profile	3rd to last word in sentence	6	10	17	29	35	43	59
Single-position word profile	4th to last word in sentence	7	11	19	31	36	45	60
Multi-position word profile	First four words in sentence	22	31	41	55	60	67	77
Multi-position word profile	First eight words in sentence	19	27	38	51	57	63	75
Multi-position word profile	Last four words in sentence	10	15	24	37	43	51	65
Multi-position word profile	Last eight words in sentence	11	16	25	38	43	52	65
Collocation profile	two words	17	24	34	48	54	61	74
Collocation profile	three words	3	6	11	21	27	35	53

The collocation algorithms performed poorly as well. In the case of the 3-word collocation algorithm—which is the least successful of all the algorithms tested in this study—this lack of success is not surprising, as *one of the* is the only three-word collocation that is frequent enough (i.e. it occurs in at least two of each author's forty texts) to be included in the profiles. On the other hand, the failure of the 2-word collocation algorithm was unexpected: even though it contains 102 collocations, and even though individual words have proven to be good indicators of authorship, the 2-word collocation algorithm does not even achieve 75% accuracy when distinguishing between two authors.

Positional stylometry has often been criticized in the past, and based on these results it would appear that the critics have been justified: positional stylometry measurements have proven to be poor indicators of authorship.

4.7 N-gram frequency

Table 8 presents the results of testing the attribution algorithms based on the relative frequency of character-level n-grams.

The *n*-gram algorithms are some of the most accurate techniques tested in this study. The most accurate *n*-gram algorithms are those based on the frequency of sequences of two and three characters: the 2- and 3-gram algorithms can distinguish between two possible authors with 94% accuracy, and can distinguish successfully between up to ten possible authors. Overall, the 2-gram algorithms are slightly more successful, barely outperforming the 3-gram algorithms when distinguishing between larger sets of possible authors. From here the performance of the algorithms steadily falls off. Interestingly, the size of the shorter *n*-gram profiles seems to matter very little, although the most successful 2-, 3-, 4- and 5-gram algorithms are based on the 10-limit profiles. On the other hand, the size of the longer n-grams profiles is significant: the six-, seven-, eight- and nine-gram algorithms performed best when the size of the profiles was maximized.

These results contradict past research: Keselj *et al.* (2003), Peng *et al.* (2003), and Clement and Sharp (2003) all achieved their best results using longer *n*-grams. But there has always been good reason to question the conclusions of these researchers, because long n-grams are known to be good indicators of topic and are often used in topic-based text classification. There is no possible way that any textual measurement can be both a

Table 8 *N*-gram frequency results

Textual measurement		Test accuracy (%) Possible authors						
Type	**Limit**	**40**	**20**	**10**	**5**	**4**	**3**	**2**
2-gram profile	In at least two texts per author	58	69	77	84	86	89	94
2-gram profile	In at least ten texts per author	65	72	79	86	88	91	94
2-gram profile	In at least twenty texts per author	60	69	77	85	87	90	94
3-gram profile	In at least two texts per author	56	68	75	82	85	89	92
3-gram profile	In at least ten texts per author	61	70	78	85	88	91	94
3-gram profile	In at least twenty texts per author	61	71	77	85	88	91	94
4-gram profile	In at least two texts per author	56	64	72	81	84	88	92
4-gram profile	In at least ten texts per author	55	64	73	83	85	89	93
4-gram profile	In at least twenty texts per author	49	58	68	78	82	86	91
5-gram profile	In at least two texts per author	45	54	66	77	80	84	90
5-gram profile	In at least ten texts per author	47	55	66	76	79	84	90
5-gram profile	In at least twenty texts per author	34	43	54	67	71	78	85
6-gram profile	In at least two texts per author	35	46	57	70	73	78	86
6-gram profile	In at least ten texts per author	35	45	56	68	72	78	86
6-gram profile	In at least twenty texts per author	23	31	42	56	61	68	79
7-gram profile	In at least two texts per author	34	42	45	59	64	69	81
7-gram profile	In at least ten texts per author	19	26	38	52	57	65	75
7-gram profile	In at least twenty texts per author	12	19	29	44	49	58	71
8-gram profile	In at least two texts per author	18	24	36	50	55	62	74
8-gram profile	In at least ten texts per author	9	16	25	40	46	54	68
8-gram profile	In at least twenty texts per author	7	12	21	35	41	49	66
9-gram profile	In at least two texts per author	12	18	28	41	46	55	68
9-gram profile	In at least ten texts per author	6	11	19	32	38	46	62
9-gram profile	In at least twenty texts per author	4	8	15	28	33	42	60

good general indicator of authorship and a good general indicator of subject, because the measurement would be unable to distinguish between text written by different authors on the same subject. Of course, in an attribution study, if each possible author wrote about a unique subject—as was certainly the case in Keselj *et al.* (2003) and Peng *et al.* (2003)—then a topic-based text classification algorithm would appear to be a good author-based text classification algorithm. Because of the careful experimental design of this study, it is likely that the results obtained here are more accurate: in general, shorter n-grams are probably better indicators of authorship than longer *n*-grams.

4.8 Overall results

Table 9 presents a ranked list of textual measurements, where only the most successful variants of each basic type are listed. The most general result of this study is that some of the quantitative authorship attribution algorithms have proven to be successful, and would still be considered successful even if the arbitrary 75% accuracy was raised. This is not a trivial result: in the past, critics of quantitative authorship attribution have been justified, to some extent, in questioning this basic assumptions because, until now, our measurements have never been tested together, and never on a corpus of possible authors as large and as challenging as the corpus used here. These results show that the quantitative comparison of texts is a legitimate approach to authorship attribution.

The most successful algorithm tested in this study is based on the word and punctuation mark profile. This method has never been tested before, but on this corpus of possible authors it outperforms all other methods. The only other

Table 9 Overall results

Textual measurement (Variant)	Test accuracy (%)						
	Possible authors						
	40	20	10	5	4	3	2
Word and punctuation mark profile (5-limit)	63	72	80	87	89	92	95
2-gram profile (10-limit)	65	72	79	86	88	91	94
3-gram profile (10-limit)	61	72	78	85	88	91	94
4-gram profile (10-limit)	55	64	73	83	85	89	93
Grapheme and punctuation mark profile	50	60	70	81	84	87	93
Multiposition graph profile (first and last six in word)	49	58	68	79	82	86	92
Word profile (5-limit)	48	57	67	77	80	85	88
5-gram profile (10-limit)	47	55	66	76	79	84	90
Multiposition grapheme profile (first six in word)	43	53	64	76	79	84	90
Multiposition grapheme profile (last six in word)	42	52	63	74	79	83	90
Punctuation mark profile (by character)	34	46	58	72	76	80	89
6-gram profile (10-limit)	35	45	56	68	72	78	86
Word-internal grapheme profile	28	39	51	65	70	76	85
Single-position grapheme profile (last in word)	27	36	49	63	68	73	84
Grapheme profile	25	35	47	62	67	74	83
7-gram profile (2-limit)	34	42	45	59	64	69	81
Single-position graph profile (2nd to last in word)	23	31	43	57	63	70	81
Single-position grapheme profile (1st in word)	20	30	41	56	62	69	80
Multiposition word profile (first four in sentence)	22	31	41	55	60	67	77
Word-length profile (fifteen intervals of one character)	18	26	39	54	60	68	79
Single-position word profile (1st word in sentence)	17	30	36	50	56	64	75
8-gram profile (2-limit)	18	24	36	50	55	62	74
2-word collocation profile	17	24	34	48	54	61	74
Tuldava's *LN*	11	18	31	49	55	64	77
Sentence-length profile (twelve intervals of twenty-five characters)	12	20	31	46	53	62	74
Sentence-length profile. (ten intervals of five words)	10	17	28	44	50	59	73
9-gram profile (2-limit)	12	18	28	41	46	55	68
Type–Token ratio	8	16	27	44	51	61	75
Herdan's *C*	7	14	25	42	49	59	73
Guiraud's *R*	7	13	24	41	48	58	73
Average word-length	7	12	22	39	46	55	70
Average sentence-length (in characters)	6	12	22	39	45	53	70
Average sentence-length (in words)	6	11	21	37	44	53	69
Yule's *K* and Simpson's *D*	6	10	18	33	38	49	65

algorithms that have successfully distinguished between up to ten possible authors are based on 2- and 3-gram profiles. All three of these measurements are similar in that they are sensitive to patterns in an author's use of common words and punctuation marks. The reason that the word and punctuation mark algorithm has outperformed the *n*-gram algorithms is probably because it is a more direct measurement of these two indicators of authorship: the frequency of an n-gram is more likely to be affected by the frequency of content words and hence by the meaning of a text. For example, the frequency of the 3-gram and is mainly determined by the frequency of the function word and, but its frequency is also affected by an author's use of such content words as *england* and *landmine* and *andy*—words which are not usually good indicators of authorship. On the other hand, the word and punctuation mark profile is not affected by such thematic patterns.

Other algorithms that have proven to be capable of distinguishing between up to five possible authors are based on the multiposition grapheme profile, the grapheme and punctuation profile, the word frequency profile, and the 4-gram profile. A number of other algorithms were also found to be of more limited use. These algorithms are based on the word-internal grapheme profile, the word-initial grapheme profile, the word-final grapheme profile, the basic grapheme profile, the five-, six-, seven- and eight-gram profiles, the word-length profile, the first four words in a sentence profile, the first word in a sentence profile, and two vocabulary richness measures (Tuldava's *LN* and the Type–Token ratio). It also appears that the 2-word collocation profile and the sentence-length profile (in characters) may also be useful indicators of authorship in the occasional case of disputed authorship. Overall, there is thus a fairly large battery of textual measurements that have proven to be of useful indicators of authorship.

5 Combination of Techniques

Because the most successful attribution algorithm tested in this study is based on a combination of the word and punctuation mark profiles, it would seem that an even more successful attribution algorithm would be based on an even larger number of textual measurements. This section presents the results of a *post hoc* test of two additional attribution algorithms that are based on combination of sixteen different types of measurements. The first seven algorithms were chosen because they have achieved at least 75% accuracy when distinguishing between five possible authors (word and punctuation profile, word profile, grapheme and punctuation profile, 2-, 3- and 4-gram profiles, and multi-position grapheme profile). The remaining nine algorithms were chosen so as to include the results of a wide range of textual measurements (word-length distribution in characters, sentence-length distribution in characters, Tuldava's *LN*, Type–Token ratio, word-internal grapheme profile, punctuation profile, 5-gram profile, two-word collocation profile, and multiposition word profile).

In order to combine the results of these sixteen attribution algorithms, the sets of measurements upon which each is based cannot be combined to make one gigantic textual profile. This is because many of these sets of textual measurements are in different scales, and it would therefore be inappropriate and ineffective to use the chi-squared statistic to compare all their values simultaneously. The simplistic solution adopted here is to attribute a text by applying each algorithm individually, and by then outputting the author that most of the attribution algorithms have selected.

Two variants of this combination algorithm have been tested. These variants differ in terms of how many votes are given to each of the individual attribution algorithms: in the simple version, each algorithm is given one vote; in the weighted version, each algorithm is given a number of votes based on its individual success (word and punctuation four, grapheme and punctuation three, 2-gram three, 3-gram three, word two, 4-gram two, multiposition grapheme two, punctuation two, 5-gram two, word-length one, sentence-length one, Tuldava's *LN* one, Type–Token ratio one, word-internal grapheme one, 2-word collocation one, and multi-position word one).

Table 10 presents the results of testing these two combination algorithms. The results of the word and punctuation and the 2-gram algorithms are included as well.

The two combination algorithms are the most accurate algorithms tested in this entire study. The simple combination algorithm equaled or bettered the best individual algorithms on six out of the seven tests. It did not perform as well in the forty author test because most of the sixteen algorithms performed very poorly at this level, and so their votes overwhelmed the votes of the few more successful attribution algorithms. However, this problem can be overcome, to some extent, by weighing the votes of each of the algorithms: the weighted combination algorithm has performed significantly better than every other algorithm tested in this study on all seven of the tests.

Table 10 Combination algorithm results

Textual measurement (Variant)	Test accuracy (%)						
	Possible authors						
	40	20	10	5	4	3	2
Weighted combination	69	78	85	91	93	95	97
Simple combination	58	72	82	90	92	94	96
Word and punctuation mark profile (5-limit)	63	72	80	87	89	92	95
2-gram profile (10-limit)	65	72	79	86	88	91	94

Most notably, the weighted combination algorithm is the first algorithm that has successfully distinguished between twenty possible authors, and that has distinguished between five possible authors with over 90% accuracy. Based on these results, it would therefore appear that the best approach to quantitative authorship attribution is one that is based on the results of as many proven attribution algorithms as possible, where the significance of each individual attribution is weighted according to the individual performance of its algorithm.

However, it should be made clear that the evaluation of the two combination algorithms was an unplanned experiment: the individual algorithms were tested first, and then the most successful algorithms were combined and retested on the same dataset. For this reason it was already very likely that the combination algorithms would outperform the individual algorithms. Nonetheless, this does not weaken the strength of the conclusion, because it has only been concluded that when attempting to resolve a case of disputed authorship an investigator should apply a variety of attribution algorithms: no claim has been made that this specific combination of algorithms is the most generally applicable combination. It is responsibility of the investigator to determine which combination can best distinguish between a particular set of possible authors.

6 Conclusion

This article has presented the results of testing a wide range of attribution algorithms on a large and carefully constructed corpus of possible authors. For the first time in the history of quantitative authorship attribution, investigators now have access to reliable data about which of our textual measurements are the most useful for attributing authorship.

Based on the results of this study, the following general procedure is proposed to resolve cases of disputed authorship. First, the investigator must identify a valid set of possible authors through an analysis of the external evidence of the anonymous text. Second, the investigator must compile a corpus of possible authors by collecting a large sample of each author's writings, which are as stylistically similar as possible to the anonymous text. Third, the investigator should test a wide range of attribution algorithms on the corpus of possible authors so as to establish which algorithms can best distinguish between that particular set of possible authors. Fourth, the investigator should test various weighted combinations of the best algorithms on the same corpus of possible authors. Finally, once an acceptably accurate combination of algorithms has been identified, the investigator can then use this algorithm to compare the anonymous text to each author-based corpus, in order to determine which possible author's writing sample is the best match.

Acknowledgements

I would like to thank Douglas Biber, Paul McFetridge, Joseph Rudman, and Maite Taboada.

References

Baayen, H., van Halteren, H., and Tweedie, F. (1996). Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, **11:** 110–20.

Bennett, W. R. (1976). *Scientific and Engineering Problem-Solving with the Computer*. Englewood Cliffs, NJ: Prentice Hall, Inc.

Bloch, B. (1948). A set of postulates for phonemic analysis. *Language*, **24:** 3–46.

Brinegar, C. S. (1963). Mark Twain and the Quintus Curtius Snodgrass letters: a statistical test of authorship. *Journal of the American Statistical Association*, **58:** 85–96.

Burrows, J. F. (1992). Not unless you ask nicely: the interpretative Nexus between analysis and information. *Literary and Linguistic Computing*, **7:** 91–109.

Burrows, J. F. and Craig, H. (2001). Lucy Hutchinson and the authorship of two seventeenth-century poems: a computational approach. *The Seventeenth Century*, **16:** 259–82.

Burrows, J. F. and Hassall, A. J. (1988). *Anna Boleyn* and the authenticity of fielding's feminine narrative. *Eighteenth Century Studies*, **21:** 427–53.

Butler, C. S. (ed.) (1992). *Computers and Written Texts*. Oxford: Blackwell.

Chaski, C. E. (2001). Empirical evaluation of language-based author identification techniques. *Forensic Linguistics*, **8:** 1–65.

Clement, R. and Sharp, D. (2003). Ngram and Bayesian classification of documents. *Literary and Linguistic Computing*, **18:** 423–47.

Eddy, H. T. (1887). The characteristic curves of composition. *Science*, March 25, **1887**: 297.

Ellegård, A. (1962a). *A Statistical Method for Determining Authorship: 1769–72*. Gothenburg: Acta Universitatis Gothoburgensis.

Ellegård, A. (1962b). *Who Was Junius?* Stockholm: Almquist and Wiksell.

Forsyth, R. S. and Holmes, D. (1996). Feature-finding for text classification. *Literary and Linguistic Computing*, **11:** 163–74.

Forsyth, R. S., Holmes, D., and Tse, E. (1999). Cicero, Sigonio, and Burrows: investigating the authenticity of the *Consolatio*. *Literary and Linguistic Computing*, **14:** 375–400.

Foster, D. (1989). *'Elegy' by W.S.: A Study in Attribution*. Cranbury, NJ: Associated University Presses.

Fucks, W. (1952). On mathematical analysis of style. *Biometrika*, **39:** 122–9.

Fucks, W. (1954). On *Nahordnung* and *Fernordnung* in samples of literary texts. *Biometrika*, **41:** 116–32.

Fucks, W. and Lauter, J. (1965). Mathematische Analyze des Literarischen Stils. In Kreuzer, H. and Gunzenhausers, R. (eds), *Mathematik und Dichtung*. Munich: Nymphenburger Verlagsbuckhandlung.

Herdan, G. (1960). *Type Token Mathematics*. The Hague: Mouton & Co.

Herdan, G. (1965). Discussion of Morton 1965a. *Journal of the Royal Statistical Society A*, **128:** 229–31.

Herdan, G. (1966). *The Advanced Theory of Language as choice and Chance*. New York: Springer-Verlag.

Hockett, C. F. (1958). *A Course in Modern Linguistics*. New York: MacMillan.

Holmes, D. (1992). A stylometric analysis of mormon scripture and related texts. *Journal of the Royal Statistical Society A*, **155:** 91–120.

Holmes, D. and Forsyth, R. (1995). The *Federalist* revisited: new directions in authorship attribution. *Literary and Linguistic Computing*, **10:** 111–27.

Holmes, D., Gordon, I., and Wilson, C. (2001). A widow and her soldier: stylometry and the American civil war. *Literary and Linguistic Computing*, **16:** 403–20.

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

Kenny, A. (1978). *The Aristotelian Ethics: A Study of the Relationship between the Eudemian and Nicomachean Ethics of Aristotle*. Oxford: Clarendon Press.

Keselj, V., Peng, F., Cercone, N., and Thomas, C. (2003). N-gram based author profiles for authorship attribution. *Pacific Association for Computational Linguistics*.

Kjetsaa, G. (1978). The Battle of The Quiet Don: another pilot study. *Computers and the Humanities*, **11:** 341–6.

Ledger, G. (1995). An exploration of differences in the pauline epistles using multivariate statistical analysis. *Literary and Linguistic Computing*, **10:** 85–97.

Ledger, G. and Merriam, T. (1994). Shakespeare, Fletcher, and the two noble kinsmen. *Literary and Linguistic Computing*, **9:** 119–24.

Levison, M., Morton, A. Q., and Wake, W. C. (1966). Some statistical features of the Pauline Epistles,'. *Journal of the Royal Philosophical Society*, **3:** 129–48.

Mascol, C. (a.k.a. W. B. Smith). (1888a). Curves of pauline and pseudo-pauline style I. *Unitarian Review*, **30:** 452–60.

Mascol, C. (a.k.a. W. B. Smith). (1888b). Curves of pauline and pseudo-pauline style II. *Unitarian Review*, **30:** 539–46.

Mendenhall, T. C. (1887). The characteristic curves of composition. *Science*, **11:** 237–49.

Mendenhall, T. C. (1901). A mechanical solution to a literary problem. *Popular Science Monthly*, **9:** 97–110.

Merriam, T. (1979). What Shakespeare wrote in Henry VIII (Part I). *The Bard*, **2:** 81–94.

Merriam, T. (1980). What Shakespeare wrote in Henry VIII (Part II). *The Bard*, **2:** 111–8.

Merriam, T. (1982). The authorship of *Sir Thomas More*. *ALLC Bulletin*, **10:** 1–7.

Merriam, T. (1988). Was hand B in *Sir Thomas More* Heywood's autograph? *Notes and Queries*, **233:** 455–8.

Merriam, T. (1994). Letter frequency as a discriminator of authorship. *Notes and Queries*, **239:** 467–9.

Merriam, T. (1998). Heterogeneous authorship in early Shakespeare and the problem of Henry V. *Literary and Linguistic Computing*, **13:** 15–28.

Michaelson, S. and Morton, A. Q. (1972). The new stylometry: a one-word test of authorship for Greek writers. *The Classical Quarterly*, **22:** 89–102.

Morton, A. Q. (1965). The authorship of Greek prose. *Journal of the Royal Statistical Society A*, **128:** 169–233.

Morton, A. Q. (1978). *Literary Detection: How to Prove Authorship and Fraud in Literature and Documents*. New York: Scribners.

Morton, A. Q. and Levison, M. (1966). 'Some indicators of authorship in Greek prose. In Leed (ed.), *The Computer and Literary Style*. Kent, OH: Kent State University Press, pp. 141–79.

Morton, A. Q. and McLeman, J. (1964). *Christianity in the Computer Age*. New York: Harper & Row: Publishers.

Morton, A. Q. and McLeman, J. (1966). *Paul, The Man and the Myth*. New York: Harper and Row.

Mosteller, F. and Wallace, D. (1963). Inference in an authorship problem. *Journal of The American Statistical Association*, **58:** 275–309.

Mosteller, F. and Wallace, D. (1964). *Inference and Disputed Authorship: The Federalist*, 1st edn. Reading, MA: Addison-Wesley.

Mosteller, F. and Wallace, D. (1984). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*, 2nd edn. New York: Springer-Verlag.

O'Brien, D. P. and Darnell, A. C. (1982). *Authorship Puzzles in the History of Economics: A Statistical Approach*. London: Macmillan.

O'Donnell, B. (1966). Stephen Crane's *The O'Ruddy*: A Problem In Authorship Discrimination. In Leed (ed.), *The Computer and Literary Style*. Kent, OH: Kent State University Press, pp. 107–15.

Peng, F., Schuurmans, D., Keselj, V., and Wang, S. (2003). Language Independent Authorship Attribution Using Character Level Language Models. *Tenth Conference of the European Chapter of the Association for Computational Linguistics*. Budapest, Hungary.

Pollatschek, M. and Radday, Y. T. (1981). Vocabulary richness and concentration in Hebrew Biblical literature. *Association for Literary and Linguistic Computing Bulletin*, **8:** 217–31.

Pollatschek and Radday, Y. T. (1985). Vocabulary Richness and Concentration. In Radday, Y. T. and Shore, H. (eds), *Genesis – an Authorship Study*. Rome: Biblical Institute.

Radday, Y. T. (1970). Isaiah and the computer: a preliminary report. *Computers and the Humanities*, **5:** 65–73.

Radday, Y. T. and Shore, H. (eds) (1985). *Genesis – An Authorship Study*. Rome: Biblical Institute.

Smith, M. W. A. (1983). Recent experience and new developments of methods for the determination of authorship. *Association for Literary and Linguistic Computing Bulletin*, **11:** 73–82.

Smith, M. W. A. (1991). The authorship of the revenger's tragedy. *Notes and Queries*, **236:** 508–11.

Smith, M. W. A. (1992). The problem of acts I-II of Pericles. *Notes and Queries*, **237:** 346–55.

Smith, M. W. A. (1993). Edmund Ironside. *Notes and Queries*, **238:** 202–5.

Somers, H. and Tweedie, F. (2003). Authorship attribution and pastiche. *Computers and the Humanities*, **37:** 407–29.

Tweedie, F. and Baayen, H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, **32:** 323–53.

Tweedie, F., Holmes, D., and Corns, T. (1998). The provenance of *De Doctrina Christiana*, attributed to John Milton: a statistical investigation. *Literary and Linguistic Computing*, **13:** 77–87.

Usher, S. and Najock, D. (1982). A statistical study of authorship in the corpus lysiacum. *Computers and the Humanities*, **16:** 85–105.

Vickers, B. (2002). *Counterfeiting Shakespeare*. Cambridge University Press.

Wake, W. C. (1957). Sentence-length distributions of Greek authors. *Journal of the Royal Statistical Society A*, **120:** 331–46.

Williams, C. B. (1940). A note on the statistical analysis of sentence-length as a criterion of literary style. *Biometrika*, **31:** 363–90.

Williams, C. B. (1970). *Style and Vocabulary*. New York: Hafner Publishing Co.

Yule, G. U. (1939). On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship. *Biometrika*, **31:** 356–61.

Yule, G. U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge, UK: Cambridge University Press.