1   **Attributing the *Bixby Letter* using n-gram tracing**

2

3   Jack Grieve[1], Emily Chiang[2], Isobelle Clarke[1], Hannah Gideon[2], Annina Heini[2], Andrea Nini[3]

4   and Emily Waibel[2]

5

6   [1] University of Birmingham

7   [2] Aston University

8   [3] University of Manchester

9

15   **Abstract**

16   There is a long-standing debate around the authorship of the *Bixby Letter*, one of the most

17   famous pieces of correspondence in American history. Despite being signed by President

18   Abraham Lincoln, some historians have claimed that its true author was John Hay, Lincoln's

19   personal secretary. Analyses of the letter have been inconclusive in part because the text

20   totals only 139 words and is thus far too short to be attributed using standard methods. To

21   test whether Lincoln or Hay wrote this letter, we therefore introduce and apply a new

22   technique for attributing short texts called *n-gram tracing*. After demonstrating that our

23   method can distinguish between the known writings of Lincoln and Hay with a very high

24   degree of accuracy, we use it to attribute the *Bixby Letter*, concluding that the text was

25   authored by John Hay – rewriting this one episode in the history of the United States and

26   offering a solution to one of the most persistent problems in authorship attribution.

27

28   **Keywords:** American History, Authorship Attribution, Computational Social Science,

29   Corpus Linguistics, Forensic Linguistics, John Hay, Abraham Lincoln, Stylistics, Stylometry

30

35   **Attributing the Bixby Letter using n-gram tracing**

36

37   **1. Introduction**

38   On the 21st of November 1864, only five months before he was assassinated, Abraham

39   Lincoln, the 16th President of the United States, sent a short letter of condolence to Lydia

40   Bixby of Boston, a widow whose five sons were believed to have died in the Civil War. The

41   original letter was lost, but the Adjutant General of Massachusetts, who had requested the

42   letter from the Department of War on the widow's behalf, also sent a copy to the *Boston*

43   *Evening Transcript*, who published the letter on the 24th of November (see Table 1). The

44   *Bixby Letter* would go on to become one of America's most famous pieces of

45   correspondence, praised for its sentiment and style and counted among Lincoln's greatest

46   texts along with the *Gettysburg Address,* the *Second Inaugural Address,* and the

47   *Emancipation Proclamation*. The authorship of the letter*,* however, has long been the subject

48   of debate, with some historians arguing that its true author was John Hay – Lincoln's young

49   assistant and the future Secretary of State under William McKinley and Theodore Roosevelt.

50   **Table 1   The *Bixby Letter* (*Boston Evening Transcript,* 25 November 1864)**

EXECUTIVE MANSION,
WASHINGTON, NOV. 21, 1864.
Dear Madam,—

I have been shown in the files of the War Department a statement of the Adjutant General of Massachusetts, that you are the mother of five sons who have died gloriously on the field of battle.

I feel how weak and fruitless must be any words of mine which should attempt to beguile you from the grief of a loss so overwhelming. But I cannot refrain from tendering to you the consolation that may be found in the thanks of the Republic they died to save.

I pray that our Heavenly Father may assuage the anguish of your bereavement, and leave you only the cherished memory of the loved and lost, and the solemn pride that must be yours, to have laid so costly a sacrifice upon the altar of Freedom.

Yours, very sincerely and respectfully,

MRS. BIXBY.   A. LINCOLN.

51

52   A wide range of external evidence has been presented in favour of both Lincoln (e.g.

53   Barton, 1926; Basler, 1953; Randall & Current, 1955; Bullard, 1946, 1951; Emerson, 2006,

54   2008) and Hay (e.g. Butler, 1940; Wakefield, 1948; Burlingame, 1995, 1999). Hay is

55   generally acknowledged to have written much of Lincoln's correspondence, as this was the

56   task for which he was hired by John George Nicolay, Lincoln's other personal secretary,

57   after Lincoln had secured the Republican presidential nomination in May 1860 (Kushner,

58   1974). Furthermore, several reliable sources – including Nicholas Murray Butler, the

59   president of Columbia University, and Spencer Eddy, Hay's personal secretary later in life –

60   claimed that Hay had confided in them that he had written the letter. In addition, Hay kept

61   scrapbooks containing extensive records of his achievements, which included the *Bixby*

62   *Letter*, as well as references to many texts he had certainly written, including his 1883 novel

63   *The Bread Winners* and a series of letters sent to newspapers across the country in support

64   of Lincoln, both of which were initially published anonymously (Kushner & Hummel, 1977).

65   Alternatively, aside from the fact that the letter bears his name, perhaps the most convincing

66   evidence that Lincoln wrote the *Bixby Letter* is that Hay never publicly took credit for its

67   authorship, although he did take credit for other letters sent by the President. Hay and

68   Nicolay even attributed the letter to Lincoln in their biography of the President (1890) and

69   Hay's children said that their father never claimed authorship in private. Furthermore,

70   although Hay authored much of Lincoln's correspondence at that time, Lincoln did write

71   some letters, including letters of condolence, and he might have been especially likely to

72   have written this letter, as he had lost three sons himself. His one surviving son, Robert

73   Todd Lincoln, who was Hay's close friend, also asserted that his father had written the *Bixby*

74   *Letter* and that Hay had confirmed as much to him personally.

75   In addition to external evidence, internal evidence related to the style of the *Bixby*

76   *Letter* has been presented in support of both Lincoln and Hay. In 1943, Basler remarked on

77   the quality of the letter and its similarity to Lincoln's style (Burlingame, 1995); ten years later,

78   he included the letter in his *Collected Works of Abraham Lincoln*. Similarly, Bullard (1946)

79   argued that the letter was generally a better match for Lincoln's style than Hay's. A more

80   thorough analysis was presented by Nickell (1989), who identified several distinctive words,

81   phrases, and rhythms in the letter, for which he could only find analogues in Lincoln's

82   writings, including the use of alliteration and the word 'tender'. Nickell also argued that

83   Lincoln wrote in a more traditional and formal style, whereas the younger Hay wrote in a

84   more contemporary and informal style. For example, Nickell claimed that the use of the word

85   'beguile' in the letter is used with its traditional sense of 'diverting', as opposed to the more

86   modern sense of 'enticing', which is how Hay used the word in a letter Nickell quotes.

87   Burlingame (1999), however, who has been one of the strongest proponents of Hay's

88   authorship, found that Hay used 'beguile' at least 30 times in his writings, including in a

89   collection of unpublished letters, while he could find no record of Lincoln ever having used

90   the word. Burlingame (1995) also argued that various other words were indicative of Hay,

91   including 'gloriously', 'cherish', 'republic', and 'Heavenly Father'.

92   The stylistic evidence is far from definitive. Burlingame and others have claimed that

93   more passages in the *Bixby Letter* resemble Hay's known writings, while Nickell and others

94   have claimed that more resemble Lincoln's. Emerson (2006: 2) dismissed this type of

95   internal evidence outright, stating that 'one can find as many arguments in favour of

96   Lincoln's literary style as one can find for Hay's.' Developing objective methods for attributing

97   authorship, however, is the focus of considerable research in *stylometry* (Koppel et al., 2009;

98   Stamatatos, 2009), where questioned documents are attributed, for example, by comparing

99   the frequencies of common words or common word and character sequences in the text to

100   their frequencies in writing samples from each possible author. The *Bixby Letter* has never

101   been subjected to thorough stylometric analysis, at least in part, because it only contains

102   139 words; short texts are difficult to attribute using stylometric techniques because the

103   relative frequencies of linguistic features in a text can only be trusted to approximate their

104 values in an author's writings more generally if that text is long enough to contain numerous

105 tokens of those features. For example, the word 'beguile' occurs once in the *Bixby Letter*, but

106 we should not assume its author used this word on average about once every 139 words.

107 Similarly, the word 'by' does not occur in the letter, but we should not assume its author

108 never used this word at all.

109 The problem of text length has received considerable attention in stylometry, with

110 Stamatatos (2009: 553) calling it 'the most important' methodological issue in the field. Eder

111 (2015) conducted the most thorough assessment of the effect of questioned document

112 length in authorship attribution and recommended a minimum length of 5,000 words; this is a

113 very conservative limit, at least in part because his tests involved between 6 and 21 possible

114 authors, as opposed to the basic problem of 2 authors, which requires less data.

115 Alternatively, many studies have been able to successfully attribute texts of around 1,000

116 (e.g. Stamatatos et al., 2001; Burrows, 2002; Juola, 2006; Stamatatos, 2009) or 500 words

117 (e.g. Gamon, 2004; Grieve, 2007; Koppel, Schler & Argamon, 2011). Few studies have

118 attributed shorter texts, although some promising results have been achieved in the 200- to

119 500-word range (e.g. Forsyth & Holmes, 1996; Koppel et al., 2011), especially based on the

120 frequencies of relatively common parts-of-speech (e.g. Chaski, 2005; Hirst & Feiguina,

121 2007). The attribution of texts shorter than 200 words has received very little attention,

122 limited mostly to a small number of recent studies of Twitter data. Most notably, Layton et al.

123 (2010) were able to attribute posts based primarily on references to usernames, while

124 Schwartz et al. (2013) were able to attribute posts based on character and word sequences

125 that are used by only one author in their corpus. Although both methods worked well for

126 classifying posts that contained these features, a substantial proportion of posts resisted

127 attribution. Better results were achieved by Brocardo et al. (2013), who proposed a method

128 for short-text *authorship verification* – which involves testing whether an author wrote a text,

129 as opposed to *authorship attribution*, which involves selecting the most likely author from a

130   set of candidates, as in the case of the *Bixby Letter*. Their method is based on the number of

131   character sequences in the questioned document that also occur in the known writings of an

132   author. Crucially, all three of these studies measured the presence and absence of linguistic

133   features as opposed to their relative frequencies, whose value is limited in short texts.

134   Totalling only 139 words, the *Bixby Letter* is far too short to be attributed using

135   standard stylometric techniques. Short documents, however, are common in a forensic

136   context (Coulthard, 2004; Coulthard et al., 2017). For example, the mean length of texts

137   received by the German Federal Criminal Police Office between 2002 and 2005 was 248

138   words, with two thirds of incriminating texts containing fewer than 200 words (Ehrhardt,

139   2007). A common method for attributing texts of any length in forensic stylistics is to

140   manually identify features of interest in the questioned document and to then search for

141   those features in the possible author writing samples to see if they are used predominantly

142   by one suspect (e.g. McMenamin, 1993, 2002). This approach is based on the reasonable

143   assumption that the repetition of features across texts is evidence of shared authorship (see

144   Coulthard, 2004). Still feature selection is usually left to the judgment of the forensic linguist,

145   limiting the reliability of this approach in practice, although forensic linguists have recently

146   begun to apply more objective selection criteria (e.g. Wright 2017). Most notably, in terms of

147   short texts, Grant (2013) attributed a series of text messages in a murder investigation

148   through a systematic analysis of the occurrence of creative spellings (see also MacLeod &

149   Grant, 2012; Silva et al. 2011). Similarly, Nini (2018) measured the similarity of short letters

150   connected to the Jack the Ripper case based on shared word sequences. Once again, like

151   the stylometric research on short texts reviewed above, these studies focus on the

152   occurrence of features as opposed to their relative frequencies.

153   Because no generally applicable method for attributing short texts exists in

154   stylometry or forensic stylistics, in this paper, we attribute the *Bixby Letter* by applying a new

155   quantitative approach to short-text authorship attribution that we call *n-gram tracing*, which

156 builds on recent research in both fields. Our method involves first extracting all sequences of

157 linguistic forms (i.e. characters and words) that occur in the questioned document and then

158 finding the possible author who uses the highest percentage of these forms. In the

159 remainder of this paper, we describe our process of data collection, introduce and exemplify

160 n-gram tracing through the analysis of the *Gettysburg Address*, test the method on the

161 known writings of Abraham Lincoln and John Hay, and use the method to attribute the *Bixby*

162 *Letter,* showing that the text is far more likely to have been written by Hay. Finally, we

163 conclude this paper by considering the historical, methodological, and theoretical

164 significance of our study.

165

166 **2. Data**

167 For years, historians believed the original *Bixby Letter* was held in the collection of

168 Brasenose College in Oxford, but in 1925 an investigation by the *New York Times* revealed

169 that the College had no record of ever possessing the document (Emerson, 2006). A futile

170 search for the letter ensued, but eventually it was accepted that the original must have been

171 lost. Some historians even speculated that the letter had been destroyed by the Widow

172 Bixby – a woman of purportedly dubious character, who had in fact lost two as opposed to

173 five sons in the Civil War, and who was rumoured to have been a brothel owner and a

174 Confederate sympathiser (Burlingame, 1999). Because there is no original, different

175 versions of the letter are in circulation today. Variation between these versions is minimal –

176 often relating to punctuation and spacing, especially in the salutation and valediction as

177 opposed to the body of the letter – but there are some disagreements in the main text, most

178 notably involving 'any word of mine' vs. 'any words of mine' and 'tendering you' vs.

179 'tendering to you'. Given these inconsistencies, it is necessary to select a specific version of

180 the *Bixby Letter* to attribute. We chose to analyse the version printed in *Boston Evening*

181 *Transcript*, because it is the first known copy of the letter and because the original is

182   accessible online[1] (see Table 1). In our analysis, we focused on the main body of the letter,

183   which contains 3 paragraphs, 4 sentences, and 139 words.

184   To compile a corpus of Lincoln's writings, we downloaded a digitised version of

185   Balser's 1953 *The Collected Works of Abraham Lincoln*, which is provided online by The

186   Abraham Lincoln Association through the University of Michigan Library[2]. The collection

187   contains over 6,500 texts, including letters, bills, notes, notices, petitions, speeches,

188   receipts, and resolutions, dated between the 26th of May 1830 and the 14th of April 1865.

189   The collection is divided into 8 volumes and organised chronologically, aside from Volume 1,

190   which contains some of Lincoln's most important writings. After downloading the documents

191   individually, we inspected each by hand, as they often contain information in addition to the

192   main text, including dates, place names, notes, and annotations by the editors. Close

193   reading of these annotations also revealed that a number of texts were only co-authored or

194   signed by Lincoln. Any document for which we had any doubt that Lincoln was the primary

195   author was therefore removed from the corpus, including the *Bixby Letter*, leaving 5,601

196   documents totalling approximately 650,000 words. These documents were then semi-

197   automatically cleaned to remove text that was not part of the main body, including

198   salutations and valedictions from letters. In addition, because Hay became Lincoln's

199   personal secretary following his presidential nomination by the Republican Party on the 18th

200   of May 1860, we removed all texts from that date onward as they were potentially written by

201   Hay. The final Lincoln corpus used to attribute the *Bixby Letter* therefore only contains texts

202   written by Lincoln up to this date, totalling 1,085 texts and 400,747 words, with texts ranging

203   in length from 5 to 17,003 words and with a median length of 125 words. Notably, average

204   text length rises from around 100 words in Lincoln's complete corpus to 350 words in

205   Lincoln's early corpus because the complete corpus includes a large number of telegraphs

206   and short letters from his time in office.

---

[1] http://news.google.com/newspapers?nid=sArNgO4T4MoC&dat=18641125
[2] http://quod.lib.umich.edu/l/lincoln/

207   To compile a corpus of Hay's writings, we downloaded a digitized version of Volume

208   I[3] and II[4] of *The Life and Letters of John Hay,* edited by William Roscoe Thayer, which was

209   originally published in 1915. The collection is organised chronologically, and includes letters,

210   prose, poems, and diary entries spanning Hay's entire life. The collection does not contain a

211   copy of the *Bixby Letter.* As opposed to the Lincoln collection, where each text could be

212   downloaded individually, the Hay texts were grouped into chapters, interspersed with

213   extensive commentary from the editor, as well as extracts from texts written by other

214   authors. After downloading the chapters, we therefore carefully inspected each file by hand

215   and manually divided the text into individual documents. Documents of unclear provenance

216   or that were co-authored by others were excluded from the corpus. In addition, we obtained

217   other texts written by Hay from Project Gutenberg, including short stories[5], poems[6], a 1901

218   novel (*The Bread Winners*)[7]*,* and a 1903 collection of essays (*Castilian Days*)[8]. We divided

219   the two book-length texts into chapters. In total, the Hay corpus contains 577 texts totalling

220   261,126 words, with texts ranging in length from 9 to 8,954 words and a median of 159

221   words per text.

222

223   **3. N-gram Tracing**

224   In forensic linguistics, short texts are often attributed by manually selecting linguistic features

225   from the questioned document that appear to be relatively distinctive or rare and by then

226   searching for these forms in the writing samples of each possible author. Although this

227   method is logical and is regularly applied in casework, there are at least three potential

228   issues with its application. First, it is unclear how to select an exhaustive or at least an

229   unbiased feature set, as the debate around the style of the *Bixby Letter* illustrates: different

[3] http://archive.org/stream/lifeandlettersof007751mbp/lifeandlettersof007751mbp_djvu.txt
[4] http://archive.org/stream/lifelettersofjoh02inthay/lifelettersofjoh02inthay_djvu.txt
[5] http://www.gutenberg.org/cache/epub/11392/pg11392.txt
[6] http://www.gutenberg.org/cache/epub/6062/pg6062.txt
[7] http://www.gutenberg.org/cache/epub/16321/pg16321.txt
[8] http://www.gutenberg.org/cache/epub/7470/pg7470.txt

230   analysts can identify different sets of seemingly distinctive features and consequently come

231   to different attributions of the same questioned document. Second, it is unclear how to

232   control for variation in the amount of material in the possible author writing samples, which

233   often varies tremendously, as is the case here: if more text is available for one of the

234   possible authors, then the forms extracted from the questioned document have an increased

235   chance of being found in that author's sample regardless of authorship. Third, it is unclear

236   how to judge whether differences in the use of forms in the possible author writing samples

237   are sufficient in the aggregate to attribute the questioned document: because this approach

238   relies on the judgment of the analyst and therefore cannot be consistently or mechanically

239   applied, it is difficult to systematically evaluate the reliability of such methods.

240   Based on this general approach to forensic authorship analysis, but keeping these

241   three limitations in mind, we have developed a new method for attributing short texts in a

242   replicable manner that we refer to as *n-gram tracing*. The method takes the n-gram as its

243   unit of analysis, where an n-gram is defined a sequence of one or more linguistic forms (e.g.

244   1-grams, 2-grams) at any level of linguistic analysis (e.g. words, characters). For example,

245   n-grams of various types extracted from the first line of the *Bixby Letter* are presented in

246   Table 2. The basic idea behind n-gram tracing is to calculate the percentage of n-grams that

247   occur in a questioned document that also occur at least once in a possible author writing

248   sample. This process is repeated for each possible author and the text is then attributed to

249   the possible author whose writing sample contains the highest percentage of the n-grams

250   from the questioned document.

251

252

253

254 **Table 2   N-gram examples from the first sentence of the *Bixby Letter***

Level	Length	Example
Word	1	*i, have, been, shown, in, the, files, of, war, …, field, battle*
	2	*I have, have been, been shown, shown in, …, of battle*
	3	*I have been, have been shown, …, field of battle*
Character	1	*i, _, h, a, v, e, b, n, s, o, w, t, …, c, y*
	2	*i_, _h, ha, av, ve, e_, _b, be, …, ba, tl*
	3	*i_h, _ha, hav, ave, _be, bee, …, ttl, tle*

255

256   Our method is grounded in two key insights. The first is that we extract the complete

257 set of n-grams that occur in the questioned document, so as to obtain a broad and unbiased

258 feature set. The second is that we only consider the presence or absence of these n-grams

259 in the questioned document and the possible author writing samples, as opposed to their

260 relative frequencies, so as to avoid examining relative frequencies in a very short text.

261 Instead, we measure the percentage of the n-gram types found in the questioned document

262 that also occur at least once in equal-sized samples of texts drawn from each possible

263 author writing sample. Specifically, for each possible author, a random sample of texts is

264 analysed that is roughly equal in length to the total number of words in the possible author

265 writing sample with the fewest words. The author who uses a higher percentage of the n-

266 grams in these comparable samples – or equivalently the author that uses a larger number

267 of unique n-grams – is then selected as the most likely author of the questioned document.

268   To summarise, our algorithm for conducting a basic n-gram tracing analysis for

269 authorship attribution involves the following four steps:

270   1. Extract all n-grams of a particular length and level from the questioned document.

271   2. Take a random sample of texts of equal size from each possible author writing

272   sample.

273   3. Measure the percentage of n-gram types found in the questioned document that

274   also occur at least once in each possible author writing sample.

275   4.   Attribute the questioned document to the possible author who uses the highest

276   percentage of these n-grams.

277   In general, n-gram tracing should be run across as many different types of n-grams as

278   possible, including both word and character-level n-grams up to a length where only a small

279   number of n-grams are occurring in the possible author writing samples. In addition, the

280   analysis can be repeated for different random samples of texts, allowing for the average

281   percentages of n-grams seen to be calculated and compared.

282   More formally, n-gram tracing involves measuring and comparing the similarity

283   between the set of n-grams occurring in a questioned document and the set of n-grams

284   occurring in each possible author writing sample. Specifically, we use the *Overlap*

285   *Coefficient* (Vijaymeena & Kavitha, 2016; Oakes, 2014), which measures the similarity

286   between two sets (X, Y) by dividing size of the intersection of those two sets (i.e. the number

287   of shared elements) by the size of the smaller set (i.e. the total number of elements):

$$\frac{|X \cap Y|}{min(|X|, |Y|)}$$

288   In the context of n-gram tracing, this amounts to dividing the number of linguistic features, in

289   our case n-grams, shared by the questioned document (Q) and a possible author writing

290   sample (A) by the number of features in the questioned document, which should always be

291   considerably smaller than in the number of features in the possible author writing sample.

$$\frac{|Q \cap A|}{|Q|}$$

292   This process is then repeated for all possible authors, using comparable writing samples,

293   and the questioned document is then attributed to the possible author with the highest

294   Overlap Coefficient.

295   Although the Overlap Coefficient is rarely used in stylometry (although see Brocardo

296   et al., 2013), the closely related *Jaccard Index*, which uses the size of the union of the two

297   sets as the denominator as opposed to the size of the smaller set, has been applied in

298   numerous recent authorship studies especially by forensic linguists (e.g. Grant, 2013;

299   Wright, 2017; Nini, 2018). We prefer the Overlap Coefficient primarily because it provides a

300   more meaningful metric of stylistic difference, directly measuring the percentage of the

301   features in the questioned document that also occur in the possible author writing sample.

302   Alternatively, the Jaccard Index measures the percentage of features shared by the

303   questioned document and the possible author writing sample, which is less interpretable, as

304   writing samples are usually far longer than questioned documents.

305   The results of n-gram tracing can also be visualised by calculating the cumulative

306   percentage of n-grams seen as texts are drawn at random from each possible author's

307   writing sample and by plotting these percentages against the total number of words in these

308   texts. In this way, it is possible to graph how the percentage of n-grams seen increases for

309   each possible author as the amount of data seen increases. To ensure the results are not

310   dependent on the random sampling of texts, this analysis can be repeated several times on

311   many different random sequences of texts and the average cumulative percentages of n-

312   grams seen can then be calculated and plotted at regular intervals of total words seen (e.g.

313   up to 5,000 words, up to 10,000 words, etc.). In general, these traces will rise rapidly at first

314   and often overlap, but as more texts are analysed, the traces will flatten out, as fewer new n-

315   grams will be encountered (see Zipf, 1935), and a clear and consistent distinction between

316   the authors should become apparent. In essence, the basic n-gram tracing algorithm

317   described above involves comparing the traces for each of the possible authors at the point

318   when the curve for the author with the smallest writing sample is exhausted; however,

319   plotting these values across sample sizes provides additional information about the use of

320   the set of n-grams in the possible author corpora. Most important, inspecting these graphs

321   allows for the definitiveness of the attribution to be judged, both by comparing the degree of

322   difference between the possible authors and the consistency of the analysis as more data is

323   analysed.

324   Although n-gram tracing was inspired by the qualitative approach to authorship

325   analysis commonly applied in forensic linguistic casework, it also builds on recent

326   quantitative research in stylometry and forensic linguistics. The multivariate analysis of word

327   and character n-grams, as broadly defined here, is the standard approach in stylometry (e.g.

328   Kešelj et al., 2003; Grieve, 2007; Luyckx & Daelemans, 2008), but the more distinctive

329   aspect of our approach is that we only consider the presence and absence of these features

330   rather than their relative frequencies. A similar approach has been taken in a small number

331   of recent studies (e.g. Brocardo et al., 2013; Grant, 2013; Schwartz et al., 2013; Wright,

332   2017; Nini, 2018). Our method is most similar to the approach for short-text authorship

333   verification proposed in Brocado et al. (2013), which is based on the analysis of the

334   occurrence of all 3-5 alphabetic character n-grams in the questioned document using the

335   Overlap Coefficient. The main difference between these two techniques are that our method

336   is designed for attribution as opposed to verification and is based on a much larger and

337   more principled feature set, including both word and character-level n-grams. Our method is

338   also similar to the approach for authorship attribution proposed in Wright (2017), where the

339   occurrence of all 2-6 word n-grams in the questioned document and the possible author

340   writing samples are compared using the Jaccard Index (see also Johnson & Wright, 2014).

341   The main differences between these two techniques are that our method is designed

342   especially for short texts, controls for the size of the possible author writing sample, is based

343   on the Overlap Coefficient as opposed to the Jaccard Index, and is based on a much larger

344   feature space. In addition, our approach to visualisation is entirely new.

345

346   **4. Demonstration:** *Gettysburg Address*

347   To illustrate how n-gram tracing works, we present an analysis of the *Gettysburg Address*,

348   which was delivered by Abraham Lincoln on the 19th of November 1863 at the site of the

349   bloodiest battle of the Civil War. We selected this text because it is one of Lincoln's most

350   famous texts, drafts prove it was written by Lincoln, and it is a relatively short text (272

351   words) that postdates May 1860, like the *Bixby Letter*. There are five final versions of the

352   *Gettysburg Address* written in Lincoln's hand, which differ slightly from each other. In this

353   case, we chose to analyse the *Bliss Copy*, as it is generally considered the standard – the

354   only version signed and dated by Lincoln and the version etched into the Lincoln Memorial.

355   We then compared the *Gettysburg Address* to the texts in our Hay and Lincoln corpora using

356   a series of n-gram tracing analyses.

357   We began by extracting all 2-word n-grams from the *Gettysburg Address*, of which

358   there are 239 distinct types when we ignore case and punctuation and prohibit n-grams from

359   spanning sentences. For example, the first 2-word n-gram in the *Address* is 'four score',

360   while the last is 'the earth'. We then measured the percentage of these 2-word n-grams in

361   the complete Hay corpus (261,126 total words) and in a random sample of texts drawn from

362   the Lincoln corpus totalling 260,954 words. We found that Hay used 55% of the n-grams,

363   whereas Lincoln used 60% (64% of the n-grams occur in Lincoln's complete 400,747 word

364   corpus). Because the 2-word n-gram overlap with the Lincoln corpus is greater, this analysis

365   correctly attributes the *Gettysburg Address* to Lincoln. We also repeated the 2-word n-gram

366   tracing analysis for Lincoln with 50 different random samples of his texts, which agreed with

367   our first analysis, with a mean percentage of n-grams seen at 260,000 words of 60%.

368   To visualise the 2-word n-gram analysis, we first extracted a random sequence of

369   texts from each possible author corpus and computed the cumulative percentage of the 239

370   2-word n-grams that had been seen as each additional text was added to the analysis. We

371   then plotted these cumulative percentages of n-grams seen against the total number of

372   words seen, as presented in Figure 1. The figure contains two traces: the longer line on top

373   plots the percentage of the 239 n-grams seen for Lincoln, which reaches 64% at 400,000

374   words, while the shorter line below plots the same value for Hay, which reaches 55% at

375   260,000 words. Individual texts are marked with a cross. Notably, both traces are monotonic

376   because adding new texts can only result in new n-grams being seen. Furthermore, both

377   traces show plateaus because at times numerous texts are added to the analyses that do

378   not contain any new n-grams. As the basic analysis found, the trace for Lincoln is higher at

379   the point where Hay's trace ends around 260,000 words, but the visualisation offers further

380   support for this attribution by showing that there is a clear and consistent difference in the

381   percentage of n-grams used by the two authors after approximately 100,000 words from

382   each had been seen.

383   We also extracted 50 random sequences of texts for each author and plotted the

384   cumulative percentage of the 239 2-word n-grams that were seen as each additional text

385   was added to the analysis. All 100 traces are presented together in Figure 2 in the same

386   way as Figure 1, except that marks for individual texts have been omitted for clarity.

387   Although each trace takes a different path, Lincoln always outstrip Hay over time, confirming

388   that the attribution does not depend substantially on the randomisation procedure. In

389   addition to presenting 100 traces on the same graph, we reduced the 50 traces for each

390   author to a single aggregated trace by taking the average cumulative percentage of n-grams

391   seen across all analyses every 5,000 words. The results of this analysis are presented in the

392   second cell of Figure 3, which shows the same overall pattern as Figures 1 and 2, with

393   Lincoln once again clearly using a higher percentage of the 2-word n-grams in the

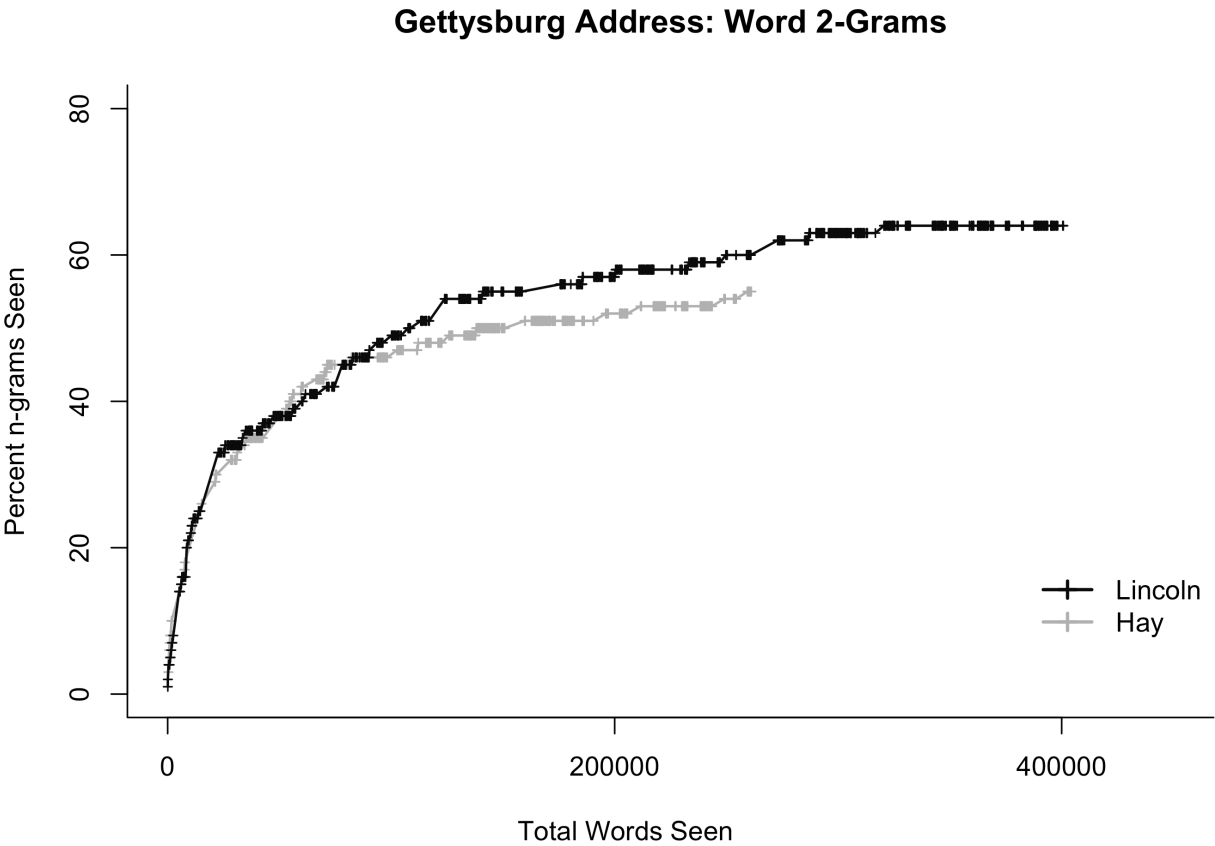394   *Gettysburg Address* than Hay.

395   In addition to 2-word n-grams, we also analysed 1-, 3- and 4-word n-grams, based

396   on the average percentage of n-grams seen in 50 random 260,000-word samples of texts.

397   The analysis was only run up to 4-word n-grams because from this point onward the Hay

398   corpus contains none of the n-grams found in the *Gettysburg Address*. The 3- and 4-word n-

399   gram analyses also correctly attributed the *Gettysburg Address* to Lincoln: 18% of 3-grams

400   for Lincoln vs. 14% for Hay and 2% of 4-grams for Lincoln vs. 0% for Hay. The 1-word n-

401   gram analysis, however, incorrectly attributed the *Gettysburg Address* to Hay. Figure 3

402   presents the aggregated n-gram traces for all analyses. Notably, the 2-, 3- and 4-word n-

403   gram analyses, which correctly attributed the document to Lincoln, appear to be far more

404   definitive than the incorrect 1-word n-gram analysis.

405   Finally, we analysed 1- to 20-character n-grams, where an n-gram could be

406   composed of any case-insensitive sequence of characters, including not only letters and

407   numbers, but punctuation marks and spaces, allowing word boundaries to be preserved,

408   although once again we did not allow n-grams to span sentences. This analysis was run for

409   n-grams of up to 20 characters in length because after this point the Hay corpus contains

410   none of the n-grams found in the *Gettysburg Address*. From 3-character n-grams onward the

411   analysis correctly attributes the document to Lincoln; the 1- and 2-character n-gram

412   analyses were inconclusive as both authors use 100% of these n-grams by 260,000 words.

413   The first 15 analyses are visualised in Figure 4, showing that the attribution becomes

414   especially clear from 7-characters onward and that the 1- and 2-character analyses both

415   reach 100% of n-grams seen almost immediately.

416   N-gram tracing therefore correctly identifies Lincoln as the author of the *Gettysburg*

417   *Address*. Overall, 21 of the 24 analyses we ran attributed the document to Lincoln, while in 2

418   of the remaining 3 cases, the analysis is inconclusive. The only analysis that incorrectly

419   attributes the *Address* to Hay is based on 1-word n-grams. To assess the degree to which

420   such misattributions affect the ability of n-gram tracing to distinguish between Lincoln and

421   Hay, we conducted a systematic evaluation of the method on the known writings of these
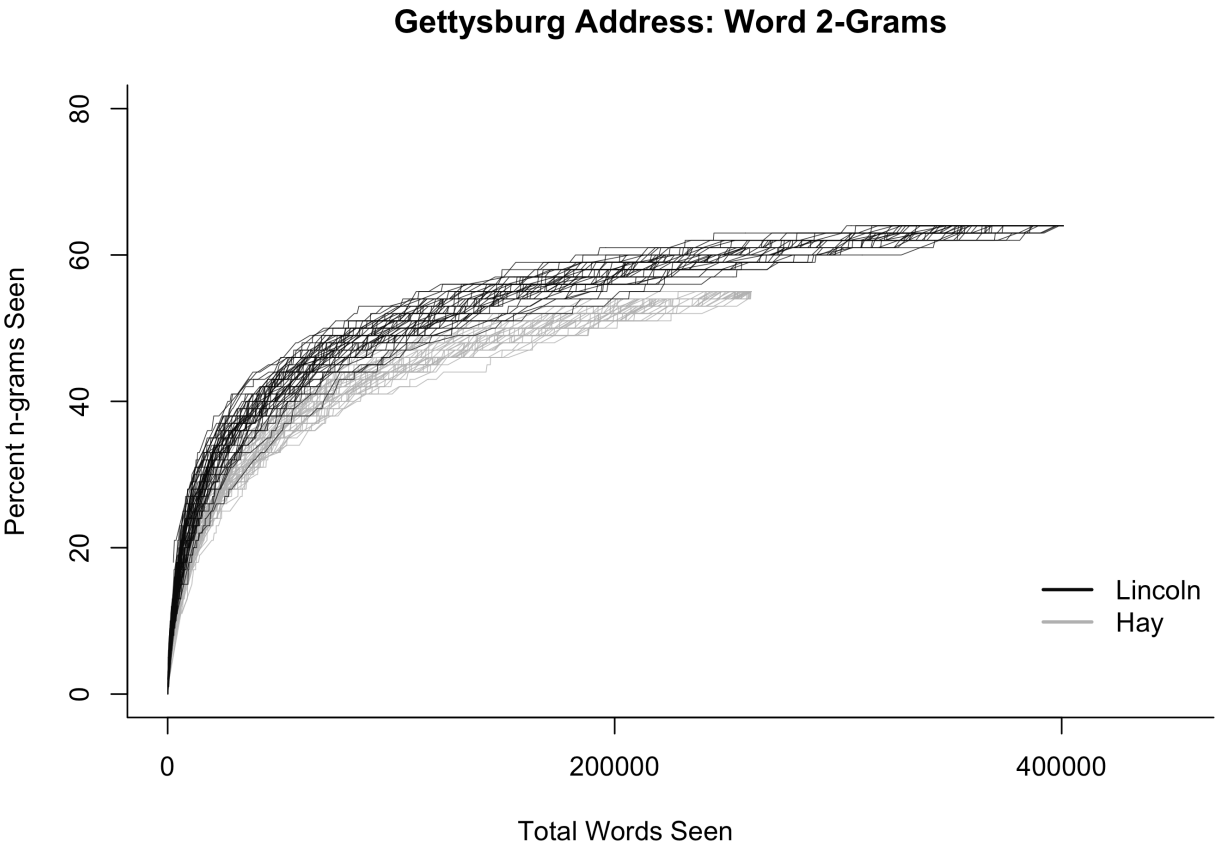
422   two authors.

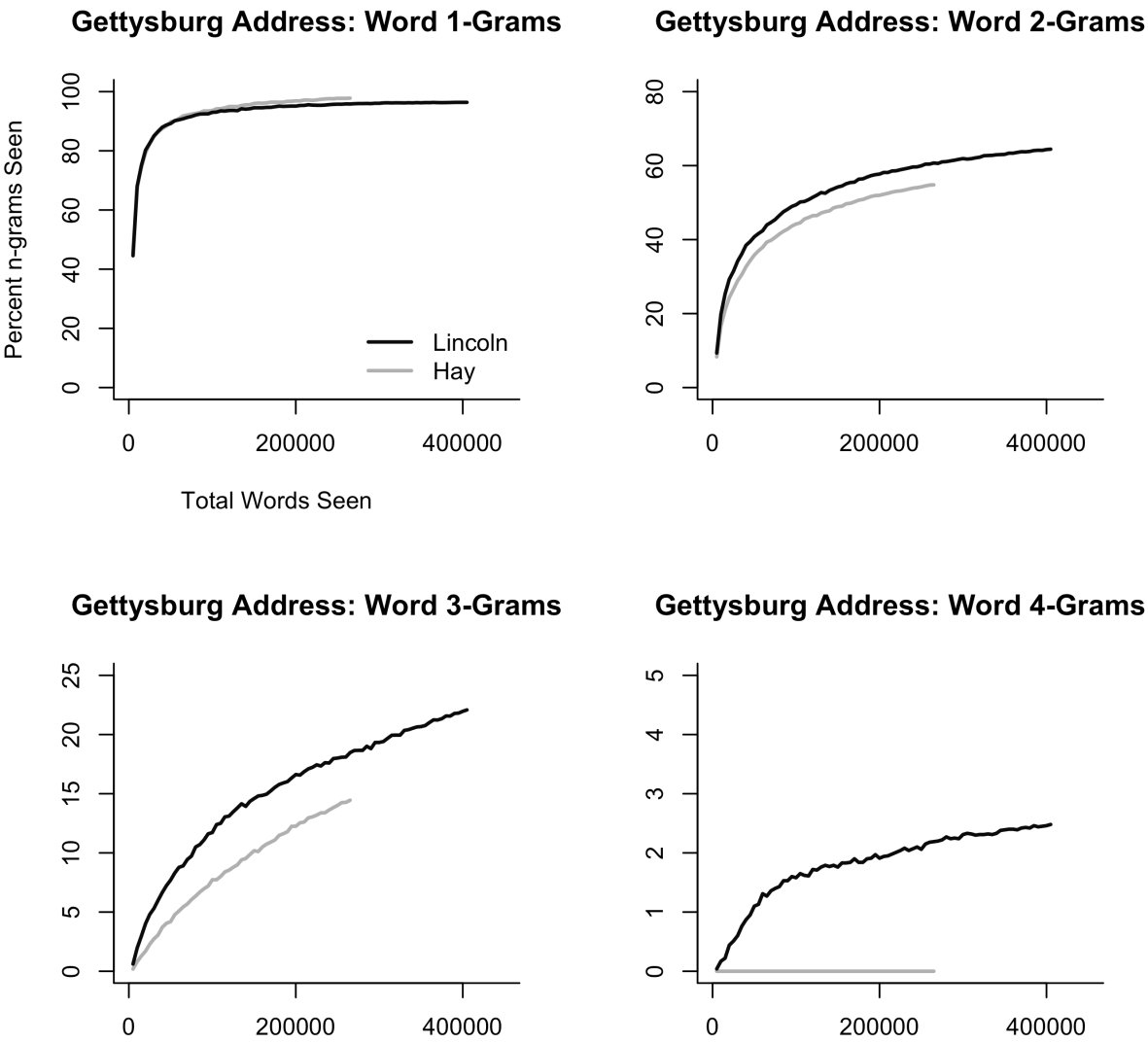423   **Figure 1**   One *Gettysburg Address* 2-word n-gram traces

424

425

426   **Figure 2**   Fifty *Gettysburg Address* 2-word n-gram traces

Gettysburg Address: Word 2-Grams
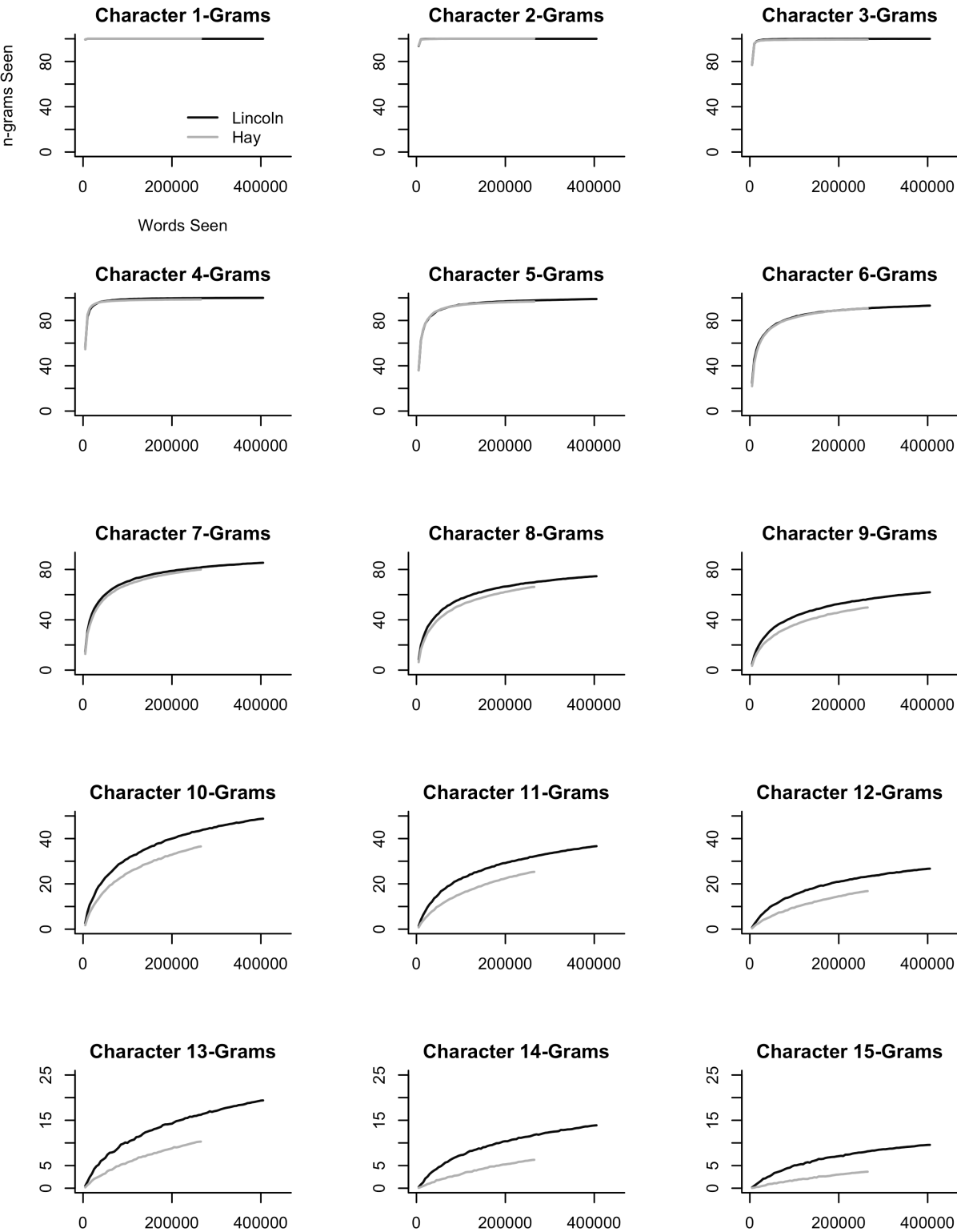
427

428

429 **Figure 3** *Gettysburg Address* word-level aggregated n-gram traces

430

431

432 **Figure 4** *Gettysburg Address* character-level aggregated n-gram traces

433

434

435

436 ## 5. Evaluation

437 Before any method for authorship attribution can be used to resolve a case of disputed

438 authorship, it must be shown that the method can distinguish between the writings of the

439 possible authors under consideration with a reasonable degree of accuracy. If the method

440 can correctly classify the known writings of those authors, then it can be used to attribute the

441 questioned document, assuming its true author is one of the authors under consideration.

442 This is the approach taken here: in this section, we show that n-gram tracing is capable of

443 distinguishing between the writings of Lincoln and Hay with a very high degree of accuracy;

444 in the next section, we use n-gram tracing to attribute the *Bixby Letter*. We do not assess or

445 assume the general applicability of n-gram tracing. This is the subject of future research, but

446 it is not a prerequisite for the application of a method to a specific case of disputed

447 authorship (see Grant 2013).

448 To evaluate the suitability of n-gram tracing for attributing the *Bixby Letter*, we used

449 our method to attribute each text in our corpus of possible authors following a leave-one-out

450 approach to cross-validation (Zhang & Yang, 2015). In other words, we removed each of the

451 1,662 texts from our corpus one at a time (1,085 for Lincoln, 577 for Hay), and then

452 attributed that text by comparing it to the remaining texts in the corpus using n-gram tracing.

453 For each text, we compared 25 different n-gram types, including 1- to 5-word and 1- to 20-

454 character n-grams, aggregating each analysis over 10 randomised sequences of texts per

455 author, selecting the author who used the higher percentage of n-grams at 260,000 words.

456 We measured the accuracy of our attributions in various ways. For each n-gram type

457 and for each author, we calculated both the *recall* (i.e. the percentage of texts written by that

458 author that were attributed to him) and the *precision* (i.e. the percentage of texts attributed to

459 that author that were written by him), in addition to a summary F_1 score, which is essentially

460 an average of precision and recall. For each n-gram type, we also calculated the percentage

461 of texts attributed correctly across the entire analysis, although this overall measure of

462   accuracy is imbalanced, as there are nearly twice as many Lincoln texts than Hay texts in

463   the corpus. Across all analyses, we counted ties, where Lincoln and Hay had the same

464   percentage of n-grams seen at 260,000 words (often 0% or 100%), as incorrect attributions

465   for both authors. In addition, we measured the accuracy of two aggregated analyses, where

466   we selected the author returned by the majority of a series of the best performing word- and

467   character-level analyses.

468   We found tracing character-level n-grams to be an especially good way to attribute

469   the writings of Lincoln and Hay (Table 3). Overall, all analyses based on between 5- and 10-

470   grams achieved F_1 scores ≥ 0.95 for both authors, with the best results obtained using 7-

471   and 8-grams. In addition, when we selected the author chosen by a majority of the analyses

472   based on between 4- and 10-grams (i.e. the author returned by at least 4 of these 7

473   analyses), we correctly identified the author of all 1,662 texts. These results clearly attest to

474   the power of n-gram tracing for distinguishing between this set of possible authors and are

475   especially remarkable given the brevity of many of the texts, a majority of which contain

476   fewer than 200 words and 10% of which contain no more than 50 words.

477   We also found tracing word-level n-grams to be good way to attribute the writings of

478   Lincoln and Hay (Table 4), although it was not as accurate as the character-level analysis.

479   Overall, analyses based on between 1- and 3-grams achieved F_1 scores ≥ 0.90 for both

480   authors, with the best results obtained using 2-grams. In addition, when we selected the

481   author chosen by a majority of the analyses based on between 1- and 3-word n-grams (i.e.

482   the author returned by at least 2 of these 3 analyses), we achieved F_1 scores ≥ 0.95 for both

483   authors.

484 **Table 3** **Character n-gram Evaluation results**

n	Hay			Lincoln			Acc
	Rec	Pre	F_1	Rec	Pre	F_1	
1	.43	.96	.59	.12	.99	.21	.23
2	.62	.93	.74	.56	.95	.70	.58
3	.93	.86	.89	.80	.98	.88	.85
4	.98	.91	.94	.93	.99	.96	.95
5	.99	.91	.95	.94	1	.97	.96
6	.99	.93	.96	.96	.99	.97	.97
7	.97	.96	.96	.98	.98	.98	.98
8	.95	.98	.96	.99	.98	.98	.98
9	.94	.98	.96	.99	.97	.98	.97
10	.92	.99	.95	.99	.96	.97	.97
11	.91	.98	.94	.99	.95	.97	.96
12	.89	.98	.93	.99	.94	.96	.96
13	.86	.98	.92	.99	.93	.96	.94
14	.83	.97	.89	.99	.92	.95	.93
15	.79	.97	.87	.99	.90	.94	.92
16	.77	.97	.86	.98	.90	.94	.91
17	.72	.97	.83	.98	.88	.93	.89
18	.68	.95	.79	.96	.89	.92	.86
19	.63	.92	.75	.94	.88	.91	.83
20	.58	.90	.71	.92	.88	.90	.80
4-10	1	1	1	1	1	1	1

485

486

487 **Table 4** **Word n-gram Evaluation results**

n	Hay			Lincoln			Acc
	Rec	Pre	F_1	Rec	Pre	F_1	
1	.96	.91	.93	.93	.98	.95	.94
2	.91	.97	.94	.99	.96	.97	.96
3	.85	.97	.91	.98	.93	.95	.93
4	.69	.94	.80	.94	.90	.92	.85
5	.41	.83	.55	.82	.89	.85	.68
1-3	.93	.98	.95	.99	.97	.98	.97

488

489

490   In addition to identifying the most reliable n-gram types upon which to base our

491   attribution of the *Bixby Letter*, it is important to consider why our analyses of other n-gram

492   types were less accurate. Analyses based on 1- and 2-character n-grams are problematic

493   because these features are far too common in the corpus of possible authors, resulting in a

494   large number of 100% ties, as reflected by the low recall scores for both authors. We

495   therefore excluded 1- and 2- character n-grams from our main analysis of the *Bixby Letter*.

496   Alternatively, analyses based on the longest word and character n-grams are problematic

497   because these features are far too uncommon in the corpus of possible authors. For

498   example, it is entirely possible that only one 5-word n-gram in a questioned document will

499   reoccur anywhere in the corpus of possible authors; in such cases, the attribution will be

500   driven entirely by this one text, potentially leading to unreliable results. We therefore

501   restricted our main analysis of the *Bixby Letter* to n-gram types where at least 5% of the n-

502   grams found in the letter are also found in the writings of Lincoln or Hay

503   We also considered how the performance of n-gram tracing was affected by text

504   length by comparing the length of texts that were successfully and unsuccessfully attributed

505   by each analysis using a series of Wilcoxon signed-rank tests. All n-gram tracing analyses

506   for each author were found to be less successful on shorter texts ($p < 0.001$). For example,

507   the median length of Hay's texts that were successfully attributed by the 7-character n-gram

508   analysis was 160 words, whereas the median length of texts that were unsuccessfully

509   attributed was 115 words. Similarly, the median length of Lincoln's texts that were

510   successfully attributed was 127 words, whereas the median length of texts that were

511   unsuccessfully attributed was 70 words. Despite these differences, n-gram tracing still

512   attributes very short texts written by Lincoln and Hay with a very high degree of accuracy, as

513   our evaluation has shown. For example, attributing texts containing fewer than 100 words

514   using a 7-character n-gram analysis still achieves 0.94 recall for Hay (vs. 0.98 recall for

515   Hay's texts that contain 100 words or more) and 0.96 recall for Lincoln (vs. 0.99 recall for

516   Lincoln's texts that contain 100 words or more). Furthermore, by this standard, the *Bixby*

517   *Letter* is a relatively long text.

518   In summary, we found that n-gram tracing, based on a range of different n-gram

519   types, is able to distinguish between the known writings of Lincoln and Hay with a very high

520   degree of accuracy, including texts containing fewer than 100 words. We found that the

521   analysis of 4- to 12-character n-grams and 1- to 3-word n-grams was especially useful for

522   distinguishing between Lincoln and Hay. We also found that selecting the author chosen by

523   the majority of the 4- to 10-character analyses attributed all 1,662 texts in our corpus of

524   possible authors perfectly. Based on the results of our evaluation, we are therefore confident

525   using n-gram tracing to investigate whether Lincoln or Hay is more likely to have written the

526   *Bixby Letter*.

527

528   **6. Results**

529   To attribute the *Bixby Letter*, we used n-gram tracing to compare all 1- to 3-word n-grams

530   and all 3- to 16-character n-grams in the *Bixby Letter* to our Lincoln and Hay writing samples

531   based on random samples of approximately 260,000 words. Longer n-gram types were

532   excluded from our analysis because fewer than 5% of the n-grams were found to occur in

533   the Hay and Lincoln corpora. Overall, all 17 of these analyses identify Hay as the author of

534   the *Bixby Letter*. Each of these n-gram tracing analyses (excluding the 15- and 16-character

535   n-gram analyses, which are very similar to traces for the other analyses) are also visualised

536   in Figure 5, based on 50 random sequences of texts for each author, aggregated in

537   increments of 5,000 words. These traces show that clear and consistent differences

538   between Hay and Lincoln are identified by 100,000 words for all word-level analyses and for

539   all character-level analyses from 5 characters onward. The n-gram tracing analysis therefore

540   clearly attributes the *Bixby Letter* to John Hay, providing very strong stylistic evidence

541   against the standard attribution of the letter to Abraham Lincoln.

542   **Figure 5**   *Bixby Letter* **aggregated n-gram traces**

543
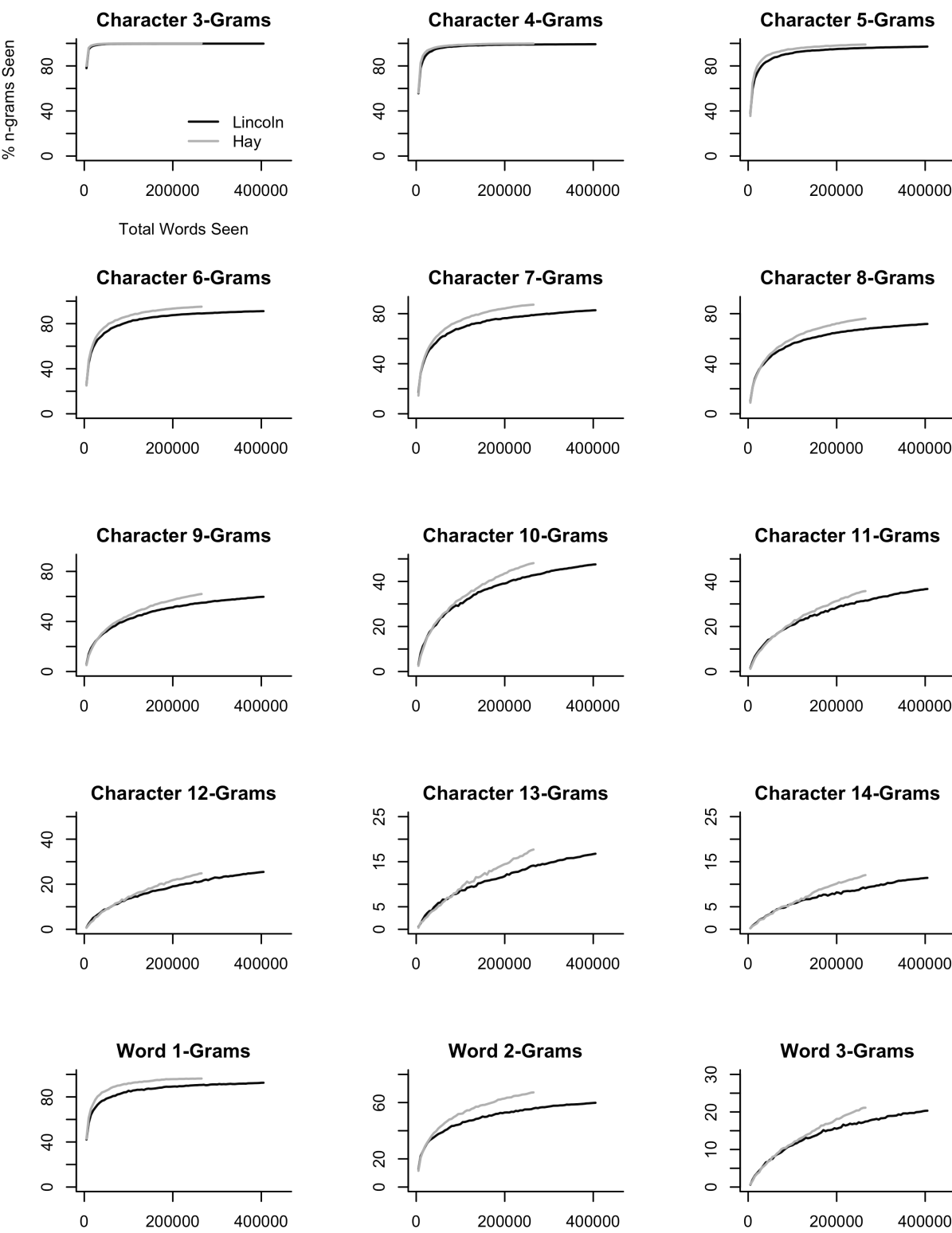
544   Although we excluded longer character n-grams from our main attribution, n-gram

545   tracing analyses based on these additional feature sets also attribute the *Bixby Letter* to

546   Hay, as does the 4-word n-gram analysis. The 5-word n-gram analysis, however, attributes

547   the *Bixby Letter* to Lincoln. This attribution is made because 'may be found in the' is the only

548   5-word n-gram out of the 115 unique 5-word n-grams in the *Bixby Letter* that occurs

549   anywhere in our corpus of possible authors, specifically in a single speech delivered by

550   Lincoln on the 11th of January 1837 at the Illinois State Assembly:

551   *If any gentleman be entitled to stock in the Bank, which he is kept out of possession*

552   *of by others, let him assert his right in the Supreme Court, and let him or his*

553   *antagonist, whichever* may be found in the *wrong, pay the costs of suit.*

554   This example illustrates the problem that arises when tracing very rare n-gram types: the

555   entire attribution can be based on a single phrase in a single text, leading to unreliable

556   results. In light of the preponderance of evidence for Hay, this one result should not diminish

557   our confidence in the attribution, especially because the meaning of 'found' in this passage

558   is different than in the *Bixby Letter*, where it means 'discovered' as opposed to 'judged'. In

559   fact, 'may be found in' is used twice by Hay, both times with the 'discovered' meaning, once

560   in an 1863 diary entry ('After every battle Lee may be found in his tent') and once in *Castilian*

561   *Days* ('This custom, more or less modified, may be found in most cities of Europe').

562   Finally, the n-grams in the *Bixby Letter* that are only used by Lincoln or Hay are

563   presented in Table 5, of which there are notably fewer for Lincoln despite being drawn from

564   a much larger corpus. Although their discriminatory value was found to be weaker, it is more

565   instructive to consider unique word-level n-grams rather than unique character-level n-

566   grams, because word-level n-grams are less common, more distinctive, and more

567   interpretable. Thematically, Hay's unique word sequences appear more evocative and

568   emotive than Lincoln's more mundane sequences – the types of constructions one might

569   expect to find in official letters sent from the Office of the President. For example, Hay's

570 unique n-grams often reference emotion (e.g. *anguish, grief*) and religion (e.g. *altar*, *pray*),

571 whereas Lincoln's often reference governmental bureaucracy (e.g. *war department, files*).

572 Grammatically, Hay's word sequences tend to contain more forms related to the construction

573 of complex noun phrases. For example, 66% of Hay's sequences contain nouns, compared

574 to 50% for Lincoln, and 49% of Hay's sequences contain determiners, compared to 32% for

575 Lincoln. Alternatively, Lincoln's word sequences tend to contain more forms related to the

576 construction of complex verb phrases. For example, 32% of Lincoln's sequences contain

577 verbs, compared to 14% for Hay, and 18% of Lincoln's sequences contain auxiliaries,

578 compared to 9% for Hay. Furthermore, 23% of Lincoln's sequences contain pronouns, while

579 only 9% of the Hay sequences do. Overall, these patterns imply that Hay's style tends to be

580 more formal than Lincoln's (see Biber 1988). Overall, while far from definitive, this closer

581 analysis of the tone and structure of the unique n-grams used by each author helps us

582 obtain a subtler understanding of the basic differences in style detected and revealed

583 through n-gram tracing.

584 Table 5   *Bixby Letter* unique word-level n-grams

n	Unique Hay n-grams	Unique Lincoln n-grams
1	*adjutant, altar, anguish, beguile, costly* (5)	*bereavement, tendering* (2)
2	*a loss, altar of, anguish of, any words, been shown, consolation that, feel how, grief of, have laid, I pray, pride that, sons who, thanks of, the altar, the anguish, the cherished, the consolation, the thanks, weak and* (19)	*a sacrifice, and fruitless, cannot refrain, father may, files of, mine which, shown in, the loved, war department, yours to* (10)
3	*and the solemn, but I cannot, from the grief, gloriously on the, thanks of the, the altar of, the anguish of, the consolation that, the grief of, the thanks of, you from the* (11)	*a statement of, and leave you, and lost and, cannot refrain from, I cannot refrain, of mine which, shown in the, statement of the, the files of, the war department* (10)

585

7. Conclusion

The historical significance of our attribution is clear. The *Bixby Letter* is one of the most famous and beautiful letters in the history of the United States and, despite on-going academic debate, it has generally been attributed to Abraham Lincoln, both by historians and the media. We have demonstrated, however, that the *Bixby Letter* was far more likely to have been authored by his 26-year-old assistant, John Hay. Assuming that only these two men could have written the *Bixby Letter*, our analysis shows that John Hay was almost certainly its primary author, providing strong linguistic support for the attributions made by Burlingame (1995, 1999) and other historians based primarily on external evidence.

Although we believe that our finding should finally lead to the official reattribution of this famous letter to John Hay, it could not detract from Abraham Lincoln's record, which was built upon far greater achievements than the *Bixby Letter*. Nevertheless, this short text is of considerable cultural, historical, and literary significance, and it is therefore important that we can now finally attribute the *Bixby Letter* with confidence to its true author. This study not only rights the historical record, but it should help historians better understand the inner workings of the Lincoln White House, arguably the most important presidency in the history of the United States. In addition, this result should remind us that John Hay was a great writer and a singular statesman, whose unwillingness to take credit for such a famous letter testifies to his humility and his love for Abraham Lincoln. Our attribution might even go some way to repairing the reputation of Mrs Lydia Bixby, for even if she was a Copperhead and a procuress, it is certainly better to have torn up a letter written by a secretary than by the President.

In addition to the historical significance of this study, the method introduced in this paper for attributing short texts represents a major step forward for authorship attribution. Short text attribution is considered to be one of the most important and difficult problems in stylometry, and n-gram tracing is a powerful solution to this problem. Our method has been

612   used here not only to attribute the *Bixby Letter,* which contains only 139 words, but over

613   1,600 texts of known authorship in both the Hay and Lincoln cannon, a majority of which are

614   shorter than 200 words and some of which are as short as 5 words. Furthermore, given that

615   n-gram tracing successfully attributed texts from various different genres without taking this

616   information into consideration, it appears that our method may also provide a solution to the

617   problem of cross-genre attribution, another fundamental challenge in stylometry and forensic

618   stylistics. Testing whether or not these types of results can be replicated over other sets of

619   possible authors is the goal of future research, in addition to testing the maximum number of

620   authors between which the method can distinguish and the minimum amount of data needed

621   for each. This is the main limitation of n-gram tracing: to reliably attribute short texts, the

622   method requires access to substantial amounts of training data for each possible author,

623   which is not always possible in historical and forensic contexts. Nevertheless, it seems clear

624   that the method could have resolved this case of disputed authorship based on far less data,

625   as many of the aggregated traces presented in Figure 5 and 6 diverge by 25,000 words.

626   More generally, the success of our method, which is rooted in forensic authorship

627   analysis, shows how insights from forensic linguistics can inform computational research on

628   authorship attribution. At the same time, our results should give forensic linguists pause.

629   This study has shown that manually selecting features, especially rare features, can lead to

630   misleading results. For example, the unique word sequences listed in Table 3 would seem to

631   be good markers of authorship, but this list, and the number of unique n-grams used by each

632   author, is only informative because it is exhaustive, especially as there are almost as many

633   unique forms for Lincoln as there are for Hay. One analyst, like Nickell, might consider the

634   word 'tendering', while another analyst, like Burlingame, might consider the word 'beguile',

635   and each will honestly come to a different conclusion, while an analyst who considers both

636   forms would come to no conclusion at all. When analysing authorship, it is therefore

637   extremely important to select a representative sample of features that is truly capable of

638   distinguishing between the authors under comparison. We have essentially taken the

639   simplest solution to this problem in this paper, attributing a text by extracting all the features

640   of a particular type that occur within it.

641   Finally, our study offers evidence in support of two theories of language use, outlined

642   in Coulthard (2004), which provide a theoretical foundation for much research in authorship

643   analysis and forensic linguistics. The first is the theory of the *uniqueness of the utterance*,

644   which claims that as sequences of words (or characters) become longer, they become less

645   likely to be repeated. This claim is supported by the results of this study, which shows that

646   the likelihood that a sequence of words or characters found in the *Bixby Letter*, or any of the

647   1,662 texts over which we evaluated our method, is repeated in the possible author writing

648   samples falls as the length of these sequences increases. In particular, n-gram tracing is

649   most successful when it focuses on n-grams of middling lengths, because sequences that

650   are too short tend to be reused by all authors, while sequences that are too long tend to be

651   reused by none. Furthermore, n-gram tracing successfully distinguishes between the

652   writings of Lincoln and Hay precisely because the likelihood of repetition falls at a slower

653   rate for the true author of these texts than for the other author. The second is the theory of

654   *idiolectal co-selection*, which states that an individual's *idiolect* – their underlying system of

655   linguistic knowledge – manifests itself during language production through the unique co-

656   selection of a variety of linguistic features. In other words, although the use of a single

657   linguistic feature is unlikely to be distinctive on its own, the co-occurrence of many features

658   will generally distinguish the linguistic output of individual authors. These co-occurrence

659   patterns are exactly the information upon which n-gram tracing is based, and our

660   unambiguous attribution of the *Bixby Letter* therefore also supports this theory of idiolectal

661   co-selection.

662   Of course, a systematic analysis of the writings of many authors and many registers

663   is needed to demonstrate that the uniqueness of the utterance and idiolectal co-selection

664   hold across the population. These are research questions we are currently pursuing, but the

665   results presented in this paper nevertheless offers initial empirical support for both of these

666   claims. Furthermore, n-gram tracing provides a replicable technique for measuring the

667   distinctiveness of linguistic forms and authorial styles. In addition to offering a solution to the

668   short text attribution problem, n-gram tracing may therefore finally provide linguists with a

669   way for judging the reality of the linguistic individual – a question of central theoretical

670   importance not only to forensic linguistics and stylometry, but many other domains of

671   linguistic inquiry.

672

673   **References**

674   **Barton, W. E.** (1926). *A Beautiful Blunder: The True Story of Lincoln's Letter to Mrs. Lydia*

675   *A. Bixby*. Indianapolis, IN: Bobbs-Merrill.

676   **Basler, R. P.** (1953). *The Collected Works of Abraham Lincoln* (8 Volumes). New

677   Brunswick, NJ: Rutgers University Press.

678   **Brocardo, M. L., Traore, I., Saad, S., and Woungang, I.** (2013). Authorship

679   verification for short messages using stylometry. In *Proceedings of the 2013*

680   *International Conference on Computer, Information and Telecommunication Systems*

681   *(CITS),* IEEE, Athens, pp. 1–6.

682   **Bullard, F. L.** (1946). *Abraham Lincoln and the Widow Bixby*. New Brunswick, NJ: Rutgers

683   University Press.

684   **Bullard, F. L.** (1951). Again, the *Bixby Letter*, *Lincoln Herald*, **37**: 26–27.

685   **Burlingame, M.** (1995). New Light on the *Bixby Letter*, *Journal of the Abraham Lincoln*

686   *Association*, **16**: 59–71.

687   **Burlingame, M.** (1999). The trouble with the *Bixby Letter*: The stirring Civil War document

688   featured in *Saving Private Ryan* grew out of a lie and probably wasn't really written by

689   Lincoln, *American Heritage*, **50**: 64-67.

690 **Burrows, J.** (2002). 'Delta': A measure of stylistic difference and a guide to likely

691 authorship, *Literary and Linguistic Computing*, **17**: 267–287.

692 **Butler, N. M.** (1940). *Across the Busy Years: Recollections and Reflections* (Volume 2).

693 New York: Charles Scribner's Sons.

694 **Chaski, C. E.** (2005). Who's at the keyboard? Authorship attribution in digital evidence

695 investigations. *International Journal of Digital Evidence*, **4**: 1–13.

696 **Coulthard, M.** (2004). Author identification, idiolect, and linguistic uniqueness, *Applied*

697 *Linguistics*, **25**: 431–447.

698 **Coulthard, M., Johnson, A., and Wright, D.** (2017). *An Introduction to Forensic*

699 *Linguistics.*

700 London: Routledge.

701 **Eder, M.** (2015). Does size matter? Authorship attribution, small samples, big problem.

702 *Digital Scholarship in the Humanities*, **30**: 167–182.

703 **Ehrhardt, S**. (2007). Forensic linguistics at the German Bundeskriminalamt. In

704 Grewendorf, G. and Rathert, M. (eds), *Formal Linguistics and Law*. Berlin: Mouton de

705 Gruyter.

706 **Emerson, J**. (2006). America's most famous letter. *American Heritage*, **57**: 1-5.

707 **Emerson, J**. (2008). New evidence from an ignored voice: Robert Todd Lincoln and the

708 authorship of *Bixby Letter*. *Lincoln Herald*, **110**: 86-116.

709 **Forsyth, R. S. and Holmes, D. I.** (1996). Feature-finding for text classification. *Literary*

710 *& Linguistic Computing*, **11**: 163-174.

711 **Gamon, M.** (2004). Linguistic correlates of style: Authorship classification with deep

712 linguistic analysis features. In *Proceedings of the 20th International Conference on*

713 *Computational Linguistics* (COLING), ACL, Geneva, Switzerland, pp. 611-617.

714 **Grant, T.** (2013). TXT 4N6: Method, consistency, and distinctiveness in the analysis of

715 SMS text messages, *Journal of Law and Policy*, **21**: 467–494.

716   **Grieve, J.** (2007). Quantitative authorship attribution: An evaluation of techniques, *Literary*

717   *and Linguistic Computing*, **22**: 251–270.

718   **Hirst, G. and Feiguina, O.** (2007) Bigrams of syntactic labels for authorship

719   discrimination of short texts. *Literary and Linguistic Computing,* **22**: 405–417.

720   **Johnson, A. and Wright, D.** (2014). Identifying Idiolect in Forensic Authorship

721   Attribution: An N-Gram Textbite Approach, *Language and Law/Linguagem E Direito*, **1**:

722   37–69.

723   **Juola, P.** (2006). Authorship attribution, *foundations and trends in information retrieval*, **1**:

724   233–334.

725   **Kešelj, V., Peng, F., Cercone, N., and Thomas, C.** (2003). N-gram-based author

726   profiles for authorship attribution. In *Proceedings of the Third Conference of the Pacific*

727   *Association for Computational Linguistics* (*PACLING 3*), Halifax, Canada, 255–264.

728   **Koppel, M., Schler, J., and Argamon, S.** (2009). Computational Methods in

729   Authorship Attribution. *JASIST*, **60**: 9–26.

730   **Koppel, M., Schler, J., and Argamon, S.** (2011). Authorship attribution in the wild.

731   *Language Resources and Evaluation*, **45**: 83–94.

732   **Kushner, H. I.** (1974). 'The Strong God Circumstance': The political career of John Hay,

733   *Journal of the Illinois State Historical Society,* **67**: 352–84.

734   **Kushner, H. I. and Hummel, S. A.** (1977). *John Milton Hay: The Union of Poetry and*

735   *Politics.* Boston, MA: Twayne Publishers.

736   **Layton, R., Watters, P., and Dazeley, R.** (2010). Authorship attribution for Twitter in

737   140 characters or less. In *Proceedings of the Second Cybercrime and Trustworthy*

738   *Computing Workshop (CTC),* Ballarat, Australia, pp. 1–8.

739   **Luyckx, K., and Daelemans, W.** (2008). Authorship attribution and verification with

740   many authors and limited data. In *Proceedings of the Twenty-Second International*

741   *Conference on Computational Linguistics (COLING 2008)*, ACL, Manchester, UK, pp.

742   513–520.

743   **MacLeod, N. and Grant, T.** (2012). Whose tweet?: Authorship analysis of micro-blogs

744   and other short form messages. In *Proceedings of the International Association of*

745   *Forensic Linguists' 10th Biennial Conference*, IAFL, Birmingham, UK, pp. 210–224.

746   **McMenamin, G. R.** (1993). *Forensic Stylistics*. Amsterdam: Elsevier.

747   **McMenamin, G. R.** (2002*). Forensic Linguistics: Advances in Forensic Stylistics*. Boca

748   Raton, FL: CRC press.

749   **Nickell, J.** (1989). Lincoln's *Bixby Letter:* A study in authorship. *Lincoln Herald,* **91**: 135–

750   140.

751   **Nini, A.** (2018). An authorship analysis of the Jack the Ripper letters. *Digital*

752   *Scholarship in the Humanities, **qx065:*** 1–16.

753   **Oakes, M. P.** (2014). *Literary Detective Work on the Computer*. Amsterdam: John

754   Benjamins Publishing Company.

755   **Randall, J. G., and Current, R. N.** (1955). *Lincoln the President*. New York: Dodd,

756   Mead.

757   **Schwartz, R., Tsur, O., Rappoport, A. and Koppel, M.** (2013). Authorship

758   Attribution of Micro-Messages. In *Proceedings of the 2013 Conference on Empirical*

759   *Methods in Natural Language Processing* (*EMNLP*), ACL, Seattle, USA, pp. 1880–

760   1891.

761   **Silva, R. S, Laboreiro, G., Sarmento, L., Grant, T., Oliveira, E. and Maia, B.**

762   (2011). 'twazn me!!! ;(' Automatic authorship analysis of micro-blogging messages. In

763   Muñoz R., Montoyo A., and Métais E. (eds), *Natural Language Processing and*

764   *Information Systems* (NLDB 2011). Berlin: Springer, pp. 161–168.

765   **Stamatatos, E.** (2009). A survey of modern authorship attribution methods, *Journal of the*

766   *American Society for Information Science and Technology*, **60**: 538–556.

767 **Stamatatos, E., Fakotakis, N., and Kokkinakis, G.** (2001) Computer-based

768   authorship attribution without lexical measures. *Computers and the Humanities,* **35**:

769   193–214.

770 **Vijaymeena, M. K., & Kavitha, K.** (2016). A survey on similarity measures in text

771   mining, *Machine Learning and Applications: An International Journal*, **3**: 19–28.

772 **Wakefield, S. D.** (1948). *Abraham Lincoln and the Bixby Letter*. New York: Wakefield, S.

773   D.

774 **Wright, D.** (2017). Using word n-grams to identify authors and idiolects. *International*

775   *Journal of Corpus Linguistics,* **22**: 212–241.

776 **Zhang, Y. and Yang, Y.** (2015). Cross-validation for selecting a model selection

777   procedure, *Journal of Econometrics*, **187**: 95–112.

778 **Zipf, G.** (1935). *The Psycho-biology of Language: An Introduction to Dynamic Philology*.

779   Boston, MA: Houghton Mifflin.