Euro. Jnl of Applied Mathematics (2021), *vol.* 32, *pp.* 1069–1105

doi:10.1017/S0956792520000157

Pull out all the stops: Textual analysis via punctuation sequences

ALEXANDRA N. M. DARMON[1], MARYA BAZZI[1,2,3], SAM D. HOWISON[1] and MASON A. PORTER[1,4]

[1]*Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK*
[2]*The Alan Turing Institute, London NW1 2DB, UK*
[3]*Warwick Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK*
[4]*Department of Mathematics, University of California, Los Angeles, Los Angeles, California 90095, USA*
emails: alexandra.darmon@hotmail.fr, mbazzi@turing.ac.uk, howison@maths.ox.ac.uk, mason@math.ucla.edu

(*Received 31 December 2018*; *revised 16 January 2020*; *accepted 12 May 2020*; *first published online 21 September 2020*)

‘‘I’m tired of wasting letters when punctuation will do, period.’’
— Steve Martin, *Twitter*, 2011

Whether enjoying the lucid prose of a favourite author or slogging through some other writer’s cumbersome, heavy-set prattle (full of parentheses, em dashes, compound adjectives, and Oxford commas), readers will notice stylistic signatures not only in word choice and grammar but also in punctuation itself. Indeed, visual sequences of punctuation from different authors produce marvellously different (and visually striking) sequences. Punctuation is a largely overlooked stylistic feature in stylometry, the quantitative analysis of written text. In this paper, we examine punctuation sequences in a corpus of literary documents and ask the following questions: Are the properties of such sequences a distinctive feature of different authors? Is it possible to distinguish literary genres based on their punctuation sequences? Do the punctuation styles of authors evolve over time? Are we on to something interesting in trying to do stylometry without words, or are we full of sound and fury (signifying nothing)?

In our investigation, we examine a large corpus of documents from Project Gutenberg (a digital library with many possible editorial influences). We extract punctuation sequences from each document in our corpus and record the number of words that separate punctuation marks. Using such information about punctuation-usage patterns, we attempt both author and genre recognition, and we also examine the evolution of punctuation usage over time. Our efforts at author recognition are particularly successful. Among the features that we consider, the one that seems to carry the most explanatory power is an empirical approximation of the joint probability of the successive occurrence of two punctuation marks. In our conclusions, we suggest several directions for future work, including the application of similar analyses for investigating translations and other types of categorical time series.

Key words: Stylometry, computational linguistics, natural language processing, digital humanities, computational methods, mathematical modelling, Markov processes, categorical time series

2020 Mathematics Subject Classification: 00A64 (Primary)

1 Introduction

"Yesterday Mr. Hall wrote that the printer's proof-reader was improving my punctuation for me, & I telegraphed orders to have him shot without giving him time to pray."

— Mark Twain, *Letter to W. Howells*, 1889

(,) . ; . , ‘ ’ , : (,) ; , ? ; , ? ?

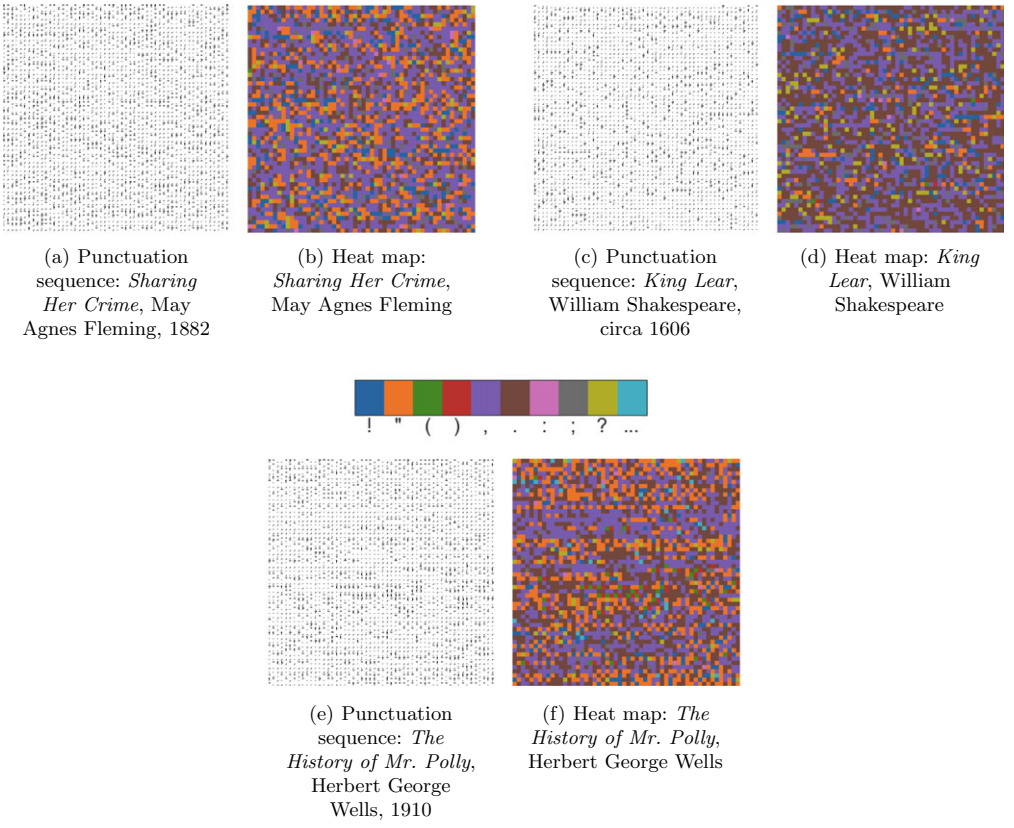
The sequence of punctuation marks above is what remains of this opening paragraph of our paper (but, to avoid recursion, without the sequence itself) after we remove all of the words. It is perhaps hard to credit that such a minimal sequence encodes any useful information at all; yet it does. In this paper, we investigate the information content of 'de-worded' documents, asking questions like the following: Do authors have identifiable punctuation styles (see Figure 1, which was inspired by the visualisations by A. J. Calhoun [3, 4]); if so, can we use them to attribute texts to authors? Do different genres of text differ in their punctuation styles; if so, how? How has punctuation usage evolved over the last few centuries?

We study sequences of punctuation marks (see Figure 1) and the number of words that separate punctuation marks. We obtain a large corpus of documents from Project Gutenberg [18]. We do not attempt to distinguish between an editor's style and an author's style for the documents in our corpus; doing so for a large corpus in an automated way is a daunting challenge, and we leave it for future efforts. We investigate whether it is possible to algorithmically assign documents to their authors, irrespective of the documents' edition(s). For ease of writing, we associate documents to authors rather than to both authors and editors throughout our paper, although we recognise that a document's writing and punctuation style can be (and usually is) a product of both.

Our paper contributes to research areas such as computational linguistics and stylometry. Broadly speaking, *computational linguistics* focuses on the development of computational approaches for processing and analysing natural language. *Stylometry*, an area of computational linguistics — and an area of cultural analytics, in the broader context of digital humanities — encompasses quantitative analysis of written text, with the goal of characterising authorship or other features [20, 35, 45]. Some of the earliest attempts at quantifying the writing style of a document include Mendenhall's work on William Shakespeare's plays in 1887 [33] and Mosteller et al.'s work on *The Federalist Papers* in 1964 [34]. The latter is often regarded as the foundation of computer-assisted stylometry (in contrast with methods that are based on human expertise) [35, 45]. Uses of stylometry include (1) authorship attribution, recognition or detection (which aim to determine whether a document was written by a given author); (2) authorship verification (which aims to determine whether a set of documents were written by the same author); (3) plagiarism detection (which aims to determine similarities between two documents); (4) authorship profiling (which aims to determine certain demographics, such as gender or other characteristics, without directly identifying an author);[1] (5) stylochronometry (which is the study

[1]For an example of 'quantitative profiling', see Neidorf et al. [36], who used stylometry to investigate stylistic features (some of which are punctuation-like, as discussed in https://arstechnica.com/science/2019/04/tolkien-was-right-scholars-conclude-beowulf-likely-the-work-of-single-author/) of *Beowulf* and concluded that it is likely the work of a single author.

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press



(a) Punctuation sequence: *Sharing Her Crime*, May Agnes Fleming, 1882

(b) Heat map: *Sharing Her Crime*, May Agnes Fleming

(c) Punctuation sequence: *King Lear*, William Shakespeare, circa 1606

(d) Heat map: *King Lear*, William Shakespeare

(e) Punctuation sequence: *The History of Mr. Polly*, Herbert George Wells, 1910

(f) Heat map: *The History of Mr. Polly*, Herbert George Wells

FIGURE 1. (a,c,e) Excerpts from ordered punctuation sequences and (b,d,f) the corresponding heat maps for books by three different authors: (a,b) May Agnes Fleming; (c,d) William Shakespeare and (e,f) Herbert George (H. G.) Wells. Each depicted punctuation sequence consists of 3000 successive punctuation marks starting from the midpoint of the full punctuation sequence of the corresponding document. The colour bar gives the mapping between punctuation marks and colours.

and detection of changes in authorial style over time) and (6) adversarial stylometry (which aims to evade authorship attribution via alteration of style).

There has been extensive work on author recognition using a wide variety of stylometric features, including lexical features (e.g., mean sentence length), syntactic features (e.g., frequencies of different punctuation marks), semantic features (e.g., synonyms) and structural features (e.g., paragraph length and number of words per paragraph). Two common stylometric features for author recognition are *n*-grams (e.g., in the form of *n* contiguous words or characters) and function words (e.g., pronouns, prepositions and auxiliary verbs). In this paper, in contrast to prior work, we focus on punctuation, rather than on words or letters. We explore several stylometric tasks through the lens of punctuation, illustrating its distinctive role in text.

According to the definition in [28], *punctuation* refers to the various systems of dots and other marks that accompany letters as part of a writing system. Punctuation is distinct from *diacritic marks*, which are typically modifications of individual letters (e.g., ç, ö and ő) and *logographs*, which are symbolic representations of lexical items (e.g., # and &). Other common symbols, such as the slash to indicate alternation (e.g., and/or) and the asterisk ‘ * ’, do not

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

fall squarely into one of these categories, but they are not considered to be true punctuation marks [28]. Common punctuation marks are the period (i.e., full stop) ‘ . ’; the comma ‘ , ’; the colon ‘ : ’; the semicolon ‘ ; ’; left and right parentheses, ‘ (’ and ‘) ’; the question mark ‘ ? ’; the exclamation point (which is also called the exclamation mark) ‘ ! ’; the hyphen ‘ - ’; the en dash ‘ – ’; the em dash ‘ — ’; opening and closing single quotation marks (i.e., inverted commas), ‘ ‘ ’ and ‘ ’ ’; opening and closing double quotation marks (which are also known as inverted commas), ‘ “ ’, and ‘ ” ’; the apostrophe ‘ ’ ’ and the ellipsis ‘ ... ’.

The aforementioned punctuation set (with minor variations) is used today in a large number of alphabetic writing systems and alphabetic languages [28]. In this sense, for a large number of languages, punctuation is a ‘supra-linguistic’ representational system. However, punctuation varies significantly across individuals, and there is no consensus on how it should be used [14, 30, 37, 39, 47]; authors, editors and typesetters can sometimes get into emphatic disagreements about it.[2] Accordingly, as a representational system, punctuation is not standardised, and it may never achieve standardisation [28].

For our study, we obtain a large corpus of documents from Project Gutenberg [18], and we extract a sequence of punctuation marks for each document in the corpus (see Section 2). Broadly, our goal is to investigate the following question: Do punctuation marks encode stylistic information about an author, a genre or a time period? (Recall that we do not distinguish between the roles of authors and editors in a document, so our use of the word ‘author’ is an expository shortcut.) Different writers have different writing styles (e.g., long versus short sentences, frequent versus sparse dialogue, and so on), and a writer’s style can also evolve over time or differ across different types of works. It is plausible that an author’s use of punctuation is — consciously or unconsciously — at least partly indicative of an idiosyncratic style, and we seek to explore the extent to which this is the case. Although there is a wealth of work that focuses on quantitative analysis of writing styles, punctuation marks and their (conscious or unconscious) stylistic footprints have largely been overlooked. Analysis of punctuation is also pertinent to ‘prosody’, the study of the tune and rhythm of speech[3] and how these features contribute to meaning [19].

To the best of our knowledge, very few researchers have explored author recognition using only punctuation-focused stylometric features [7, 17]. Additionally, the few existing works that include a punctuation-focused analysis used a very small author corpus (40 authors in [17] and 5 authors in [7]) and concentrated on the frequency with which different punctuation marks occur (ignoring, e.g., the order in which they occur). In the present paper, we investigate author recognition using features that account (partially) for both the frequency and the order of punctuation marks in a corpus of 651 authors and 14,947 documents from the Project Gutenberg database (see Section 3). Although Project Gutenberg is a popular database for the statistical analysis of language, most previous studies that used this database considered only a small number of manually selected documents [15]. Again using data from Project Gutenberg, we also employ a punctuation perspective to explore genre recognition [9, 24, 41, 42] in Section 4 and stylochronometry [6, 13, 22, 23, 38, 46, 50] in Section 5. There are not many studies of stylochronometry, and existing ones tend to be rather specific in nature (e.g., with a focus on

[2]Not that any of us would ever descend to this.

[3]An amusing illustration of prosody is the contrast between the Oxford comma, the Walken comma, and the Shatner comma. For one example, see https://www.scoopnest.com/user/JournalistsLike/529351917986934784.

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

particular authors, such as Shakespeare [50] and band members from the Beatles [23], or on particular time frames) [35,46]. Literary genre recognition (e.g., fiction and philosophy) has also received limited attention, and we are not aware of even a single study that has attempted genre recognition using punctuation. We wish to examine (1) whether punctuation is at all indicative of the style of an author, genre or time period; and if so, (2) the strength of stylistic signatures when one ignores words. In short, how much can one learn from punctuation alone?

We use machine learning for our stylometric tasks (such as author recognition) and compute several features (see Section 2.2) from each document to use as explanatory variables for these tasks. We do not seek either to try to identify the best set of features or to conduct a thorough comparison of different machine-learning methods for a given stylometric task. Instead, our goal is to give punctuation, an unsung hero of style, some overdue credit through an initial quantitative study of punctuation-focused stylometry. To do this, we examine a small number of punctuation-related stylometric features and use this set of features to investigate questions in author recognition, genre recognition and stylochronometry. To reiterate an important point, we do not account for the effects of editorial changes on an author's style, and it is important to interpret all of our findings with this caveat in mind. We offer a novel perspective on stylometry that we hope others will carry forward in their own punctuational pursuits, which include many exciting future directions.

Our paper proceeds as follows. We describe our data set (as well as our filtering and cleaning of it), punctuation-based features and classification techniques in Section 2. We compare the use of punctuation across authors in Section 3, across genres in Section 4 and over time in Section 5. We conclude and offer directions for future work in Section 6. The data set of the punctuation sequences that we use in this paper is available at https://dx.doi.org/10.5281/zenodo.3605100, and the code that we use to analyse punctuation sequences is available at https://github.com/alex-darmon/punctuation-stylometry.

2 Data and methodology

"This sentence has five words. Here are five more words. Five-word sentences are fine. But several together become monotonous. Listen to what is happening. The writing is getting boring. The sound of it drones. It's like a stuck record. The ear demands some variety. Now listen. I vary the sentence length, and I create music. Music. The writing sings. It has a pleasant rhythm, a lilt, a harmony. I use short sentences. And I use sentences of medium length. And sometimes, when I am certain the reader is rested, I will engage him with a sentence of considerable length, a sentence that burns with energy and builds with all the impetus of a crescendo, the roll of the drums, the crash of the cymbals — sounds that say listen to this, it is important."

— Gary Provost, *100 Ways to Improve Your Writing*, 1985.

2.1 Data set

We use the application-programming-interface (API) functionality of Project Gutenberg [18] to obtain our document corpus and the natural-language-processing (NLP) library SPACY [21]

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

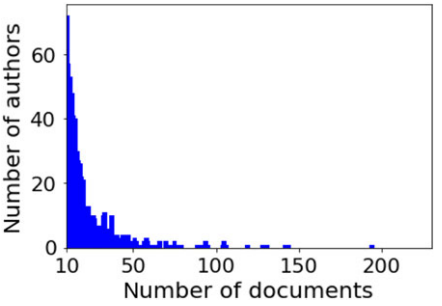


FIGURE 2. Histogram of the number of documents per author in our corpus.

to extract a punctuation sequence from each document.[4] Using data from Project Gutenberg requires several filtering and cleaning steps before it is meaningful to perform statistical analysis [15]. We describe our steps in this subsection.

We retain only documents that are written in English. (A document's language is specified in metadata.) We remove the author labels 'Various', 'Anonymous' and 'Unknown'. To try and mitigate, in an automated way, the issue of a document appearing more than once in our corpus (e.g., 'Tales and Novels of J. de La Fontaine – Complete', 'The Third Part of King Henry the Sixth', 'Henry VI, Part 3', 'The Complete Works of William Shakespeare' and 'The History of Don Quixote, Volume 1, Complete'), we ensure that any given title appears only once, and we remove all documents with the word 'complete' in the title.[5] Note that the word 'anthology' does not appear in any titles in our final corpus. We also adjust some instances in which a punctuation mark or a space appears incorrectly in Project Gutenberg's raw data (specifically, instances in which a double quotation appears as unicode or the spacing between words and punctuation marks is missing), and we remove any documents in which double quotations do not appear.[6] Among the remaining documents, we retain only authors who have written at least 10 documents in our corpus. For each of these documents, we remove headers using the Python function STRIP_HEADERS, which is available in Project Gutenberg's Python package. This yields a data set with 651 authors and 14,947 documents. We give this final list of authors in Appendix A, and we show the distribution of documents per author in Figure 2. The documents in our corpus have various metadata, such as author birth year, author death year, document 'bookshelf' (with at most one unique bookshelf per document), document subject (with potentially multiple subjects per document), document language and document rights. In some of our computational experiments, we use the following metadata: author birth year, author death year and document 'bookshelf' (which we term document 'genre', as that is what it appears to represent). Gerlach and Font-Clos [15] pointed out recently that the term 'bookshelf' may be better suited than the term 'subject' to practical purposes such as text classification, because the former encompasses broader categories and provides a unique assignment of labels to documents.

[4]Many abbreviations, such as 'etc.' and 'Mr.', are treated as words in SPACY. Therefore, SPACY does not count the periods in them as punctuation marks.

[5]It is still possible for a document to appear more than once in our corpus (e.g., 'The Third Part of King Henry the Sixth' and 'Henry VI, Part 3'). We manually remove such duplicates when investigating specific authors over time (see Section 5).

[6]Many of these documents may be legitimate ones, but we remove them to err on the side of caution.

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

For each document, we extract a sequence of the following 10 punctuation marks: the period ‘ . ’; the comma ‘ , ’; the colon ‘ : ’; the semicolon ‘ ; ’; the left parenthesis ‘ (’; the right parenthesis ‘) ’; the question mark ‘ ? ’; the exclamation mark ‘ ! ’; double quotation marks, ‘ “ ’ and ‘ ” ’ (which are not differentiated consistently in Project Gutenberg’s raw data); single quotation marks, ‘ ‘ ’ and ‘ ’ ’ (which are also not differentiated consistently in Project Gutenberg’s raw data), which we amalgamate with double quotation marks; and the ellipsis ‘ ... ’. To promote a language-independent approach to punctuation (e.g., apostrophes in French can arise as required parts of words), we do not include apostrophes in our analysis. We also do not include hyphens, en dashes or em dashes, as these are not differentiated consistently in Project Gutenberg’s raw data, and we find the choices between these marks in different documents — standard rules of language be damned — to be unreliable upon a visual inspection of some documents in our corpus. Lastly, we exclude square brackets (which are also sometimes called ‘brackets’), as they are used in metadata within the documents in Project Gutenberg.

2.2 Features

Employing standard terminology from machine learning, we use the word ‘feature’ to refer to any quantitative characteristic of a document or set of documents. We compute six features for each document k in our corpus to quantify the frequency with which punctuation marks occur, the order in which they occur, and the number of words that tend to occur between them. Specifically, we compute the following features:

(1) $\boldsymbol{f}^{1,k}$, the frequency vector for punctuation marks in a given document k;
(2) $\boldsymbol{f}^{2,k}$, an empirical approximation of the conditional probability of the successive occurrence of elements in an ordered pair of punctuation marks in document k;
(3) $\boldsymbol{f}^{3,k}$, an empirical approximation of the joint probability of the successive occurrence of elements in an ordered pair of punctuation marks in document k;
(4) $\boldsymbol{f}^{4,k}$, the frequency vector for sentence lengths in a given document k, where we consider the end of a sentence to be marked by a period, an exclamation mark, a question mark or an ellipsis;
(5) $\boldsymbol{f}^{5,k}$, the frequency vector for the number of words between successive punctuation marks in a given document k and
(6) $\boldsymbol{f}^{6,k}$, the mean number of words between successive occurrences of the elements in an ordered pair of punctuation marks in document k.

We summarise these six features in Table 1 and define each of them in the present subsection. When appropriate (and for ease of writing), we suppress the superscript k.

Let $\Theta = \{\theta_1 | \ldots | \theta_{10}\}$ denote the (unordered) set of 10 punctuation marks (see Section 2.1).[7] Let n denote the number of documents in our corpus; and let $D_k = \{\theta_1^k | \ldots | \theta_{n_k}^k\}$, with $k \in \{1, \ldots, n\}$, denote the sequence of punctuation marks in document k. The sequence D_k has n_k elements. As an example, consider the following quote by Ursula K. Le Guin (from an essay in her 2004 collection, *The Wave in the Mind*):

[7]Because there can be commas in the elements of some of the sets and sequences that we consider, we use vertical lines (instead of commas) to separate elements in sets and sequences with punctuation marks to avoid confusion.

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

Table 1. *Summary of the punctuation-sequence features that we study. See the text for details and defining formulas.*

Feature	Description	Formula
$\boldsymbol{f}^1$	Punctuation-mark frequency	(2.1)
$\boldsymbol{f}^2$	Empirical conditional probability of successive punctuation marks	(2.2)
$\boldsymbol{f}^3$	Empirical joint probability of successive punctuation marks	(2.3)
$\boldsymbol{f}^4$	Sentence-length frequency	(2.4)
$\boldsymbol{f}^5$	Frequency vector for the number for words between successive punctuation marks	(2.5)
$\boldsymbol{f}^6$	Mean number of words between successive occurrences of the elements in ordered pairs of punctuation marks	(2.7)

```
I don't have a gun and I don't have even one wife and my sentences tend
to go on and on and on, with all this syntax in them. Ernest Hemingway
would have died rather than have syntax. Or semicolons. I use a whole
lot of half-assed semicolons; there was one of them just now; that was
a semicolon after 'semicolons,' and another one after 'now.'
```

The sequence D_k for this quote is {, | . | . | . | ; | ; | ‘ | , | ’ | ‘ | . | ’}, and there are $n_k = 12$ punctuation marks. From D_k, we can calculate $\boldsymbol{f}^{1,k}$, $\boldsymbol{f}^{2,k}$ and $\boldsymbol{f}^{3,k}$.

We determine each entry of $\boldsymbol{f}^{1,k}$ from the number of times that the associated punctuation mark appears in a document, relative to the total number of punctuation marks in a document:

$$f_i^{1,k} = \frac{\left|\{\theta_l^k \in D_k \mid \theta_l^k = \theta_i\}\right|}{n_k}. \tag{2.1}$$

The feature $\boldsymbol{f}^{1,k}$ induces a discrete probability distribution on the set of punctuation marks for each document in our corpus (i.e., $\sum_{i=1}^{|\Theta|} f_i^{1,k} = 1$ for all k) and is independent of the order of the punctuation marks. For the Le Guin quote,

$$\boldsymbol{f}^1 = \begin{bmatrix} \mathbf{!} & \mathbf{"} & \mathbf{(} & \mathbf{)} & \mathbf{,} & \mathbf{.} & \mathbf{:} & \mathbf{;} & \mathbf{?} & \mathbf{...} \\ 0 & \frac{1}{3} & 0 & 0 & \frac{1}{6} & \frac{1}{3} & 0 & \frac{1}{6} & 0 & 0 \end{bmatrix},$$

where the second row indicates the elements of the vector and the first row indicates the corresponding punctuation marks. (Recall from Section 2.1 that we amalgamate opening and closing quotation marks — both double quotation marks and single quotation marks — into a single punctuation mark, so the associated entry refers to the appearance of any of those marks.) An alternative is to consider the frequency of punctuation marks relative to the number of characters or words in a document [17]. In Figure 3, we show the histograms of punctuation-mark frequencies (which are given by $\boldsymbol{f}^1$) across all documents in our corpus. These plots give an idea of the overall usage of each punctuation mark in our corpus. For instance, we see that commas and periods are (unsurprisingly) the most common punctuation marks in the corpus documents. We also observe that comma frequency varies more across documents than period frequency.

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

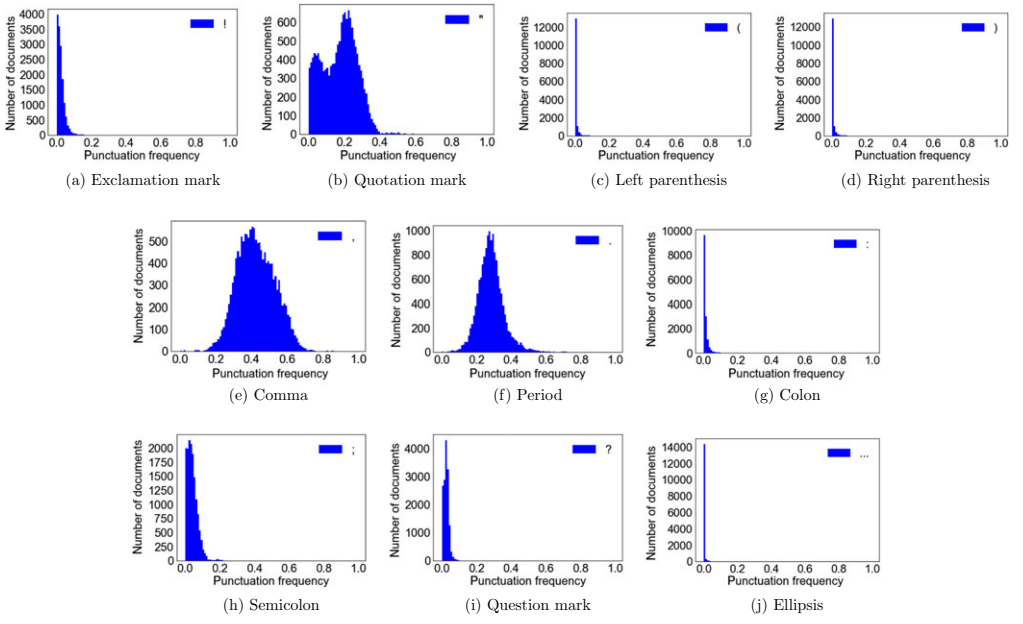


FIGURE 3. Histograms of punctuation-mark frequencies of the documents in our corpus. The horizontal axis of each panel gives the frequency of a punctuation mark binned by 0.01, and the vertical axis of each panel gives the total number of documents in our corpus with a punctuation-mark frequency in the bin. That is, the first bar of a panel for punctuation mark θ_i indicates the number of documents in our corpus for which $0 \leqslant f_i^{1,k} < 0.01$, the second bar indicates the number of documents in our corpus for which $0.01 \leqslant f_i^{1,k} < 0.02$, and so on. In descending order, the means (rounded to the third decimal) of each set $\{ f_i^{1,k} \mid k = 1, \ldots, n\}$ (which we use to construct our plot for θ_i) are 0.024 (exclamation mark), 0.175 (apostrophe), 0.006 (left parenthesis), 0.006 (right parenthesis), 0.425 (comma), 0.283 (period), 0.013 (colon), 0.041 (semicolon), 0.025 (question mark) and 0.002 (ellipsis). These numbers imply that, on average, 42.5% of the punctuation marks of a document in our corpus are commas, 28.3% are periods, 4.1% are semicolons, and so on.

Additionally, there appear to be two peaks in quotation-mark frequency: a lower peak (with a height of approximately 450 documents) at about 0.1 and a higher peak (with a height of approximately 650 documents) at about 0.25. No other punctuation mark has more than one noticeable peak; this perhaps suggests that one can cluster the documents in our corpus into two sets that are distinguished by how often they use quotation marks.

To compute $\boldsymbol{f}^{2,k}$ and $\boldsymbol{f}^{3,k}$, we consider a Markov chain on the sequence of punctuation marks and associate each punctuation mark with a state of the Markov chain. We first need two types of transition matrices. We calculate the matrix $\boldsymbol{P}^k \in [0,1]^{|\Theta|\times|\Theta|}$ from the number of times that elements in an ordered pair of punctuation marks occur successively in a document, relative to the number of times that the first punctuation mark in this pair occurs in the document:

$$P_{ij}^k = \frac{\left|\{\theta_l^k \in D_k \mid \theta_l^k = \theta_i \text{ and } \theta_{l+1}^k = \theta_j\}\right|}{\left|\{\theta_l^k \in D_k \mid \theta_l^k = \theta_i\}\right|}, \quad \text{such that} \quad \sum_j P_{ij}^k = 1. \tag{2.2}$$

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

When a punctuation mark θ_i does not appear in a document, we set all entries in the corresponding row to 0. We calculate the matrix $\widetilde{\boldsymbol{P}}^k \in [0,1]^{|\Theta|\times|\Theta|}$ from the number of times that elements in an ordered pair of successive punctuation marks occur in a document, relative to the total number of punctuation marks in the document:

$$\widetilde{P}^k_{ij} = \frac{\left|\{\theta^k_l \in D_k \mid \theta^k_l = \theta_i \text{ and } \theta^k_{l+1} = \theta_j\}\right|}{n_k}, \quad \text{such that} \quad \sum_{i,j} \widetilde{P}^k_{ij} = 1. \tag{2.3}$$

Note that $\widetilde{P}^k_{ij} = P^k_{ij} f^{1,k}_i$.

The transition matrix $\boldsymbol{P}^k$ is an estimate of the conditional probability of observing punctuation mark θ_j after punctuation mark θ_i in document k, and the transition matrix $\widetilde{\boldsymbol{P}}^k$ is an estimate of the joint probability of observing the punctuation marks θ_i and θ_j in succession in document k. The relationship $\widetilde{P}^k_{ij} = P^k_{ij} f^{1,k}_i$ ensures that rare (respectively, frequent) events are given less (respectively, more) weight in $\widetilde{\boldsymbol{P}}$ than in $\boldsymbol{P}$. For example, if an author seldom uses an ellipsis in a document, the few occurrences of it (which, arguably, are not representative of authorial style) are assigned large probabilities in $\boldsymbol{P}$ but small probabilities in $\widetilde{\boldsymbol{P}}$. For the Le Guin quote, $\boldsymbol{P}$ and $\widetilde{\boldsymbol{P}}$ are

$$\boldsymbol{P} = \begin{array}{c} \begin{array}{cccccccccc} \textbf{!} & \textbf{"} & \textbf{(} & \textbf{)} & \textbf{,} & \textbf{.} & \textbf{:} & \textbf{;} & \textbf{?} & \textbf{...} \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}, \quad \widetilde{\boldsymbol{P}} = \begin{array}{c} \begin{array}{cccccccccc} \textbf{!} & \textbf{"} & \textbf{(} & \textbf{)} & \textbf{,} & \textbf{.} & \textbf{:} & \textbf{;} & \textbf{?} & \textbf{...} \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{9} & 0 & 0 & \frac{1}{9} & \frac{1}{9} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{12} & 0 & 0 & 0 & \frac{1}{12} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{12} & 0 & 0 & 0 & \frac{1}{6} & 0 & \frac{1}{12} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{12} & 0 & 0 & 0 & 0 & 0 & \frac{1}{12} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array},$$

where the first row of each matrix indicates the corresponding punctuation mark. Observe that $\widetilde{P}_{56} < \widetilde{P}_{66}$, even though the corresponding entries equal each other in $\boldsymbol{P}$, because two successive periods occur more frequently than a period followed by a comma in Le Guin's quote.

We obtain $\boldsymbol{f}^{2,k}$ and $\boldsymbol{f}^{3,k}$ by 'flattening' (i.e., concatenating the rows of) the matrices $\boldsymbol{P}^k$ and $\widetilde{\boldsymbol{P}}^k$, respectively. For example, we obtain $\boldsymbol{f}^2$ for the Le Guin quote by appending, in order, the rows of $\boldsymbol{P}$. The feature $\boldsymbol{f}^{3,k}$ induces a joint probability distribution on the space of ordered punctuation pairs. In contrast to $\boldsymbol{f}^{1,k}$, the features $\boldsymbol{f}^{2,k}$ and $\boldsymbol{f}^{3,k}$ depend on the order in which punctuation marks occur in a document. As we will see in Section 3, the feature $\boldsymbol{f}^{3,k}$ is very effective at distinguishing different authors. We account for order with a one-step lag (i.e., each state depends only on the previous state) in $\boldsymbol{f}^{2,k}$ and $\boldsymbol{f}^{3,k}$. One can generalise these features to account for memory or 'long-range correlations' [12]. For example, the probability that a right parenthesis occurs increases after there is a left parenthesis (because a parenthetical remark requires a closing parenthesis).

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

The features $\boldsymbol{f}^{4,k}$, $\boldsymbol{f}^{5,k}$ and $\boldsymbol{f}^{6,k}$ account for the number of words that occur between punctuation marks. Let $D_k^w = \{w_0^k, w_1^k, \ldots, w_{n_k-1}^k\}$ denote the number of words that occur between successive punctuation marks in D_k, with w_0^k equal to the number of words before the first punctuation mark. Therefore, w_1^k is the number of words between punctuation marks θ_1^k and θ_2^k, and so on. The sequence D_k^w for Le Guin's comment is $\{25, 6, 9, 2, 9, 7, 5, 1, 0, 4, 1, 0\}$, where we count 'don't' as two words and we also count 'half-assed' as two words. The minimum number of words that can occur between successive punctuation marks is 0, and we cap the maximum number of words that can occur between successive punctuation marks at $n_s = 40$ and the number of words in a sentence at $n_S = 200$. Fewer than 0.05% of the sentences in our corpus exceed $n_S = 200$ words; the $n_s = 40$ cap is exceeded by fewer than 0.05% of the strings between successive punctuation marks.[8]

The entries of the feature $\boldsymbol{f}^{4,k} \in [0, 1]^{n_S \times 1}$, which quantifies the frequency of sentence lengths, are

$$f_i^{4,k} = \frac{\left|\{w_l^k \in D_k^w \mid w_l^k = i \text{ and } \theta_l, \theta_{l+1} \in \{. \mid \ldots \mid ! \mid ?\}\}\right|}{n_k}, \tag{2.4}$$

where we recall that a sentence can end in a period, an ellipsis, an exclamation mark, or a question mark. In the Le Guin quote, there are four sentences, with lengths 31, 9, 2 and 27 (in sequential order). The feature $\boldsymbol{f}^{4,k}$, an $n_S \times 1$ vector with $n_S = 200$, thus has the value $1/4$ in the 9th, 2nd, 27th and 31st positions and the value 0 in all other entries. One can also consider other measures of sentence length (e.g., the number of characters, instead of the number of words) [48].

The entries of the feature $\boldsymbol{f}^{5,k} \in [0, 1]^{n_s \times 1}$, which quantifies the frequency of the number of words between successive punctuation marks, are

$$f_i^{5,k} = \frac{\left|\{w_l^k \in D_k^w \mid w_l^k = i\}\right|}{n_k}. \tag{2.5}$$

In the Le Guin quote, recall that $D_k^w = \{25, 6, 9, 2, 9, 7, 5, 1, 0, 4, 1, 0\}$ (which includes nine unique integers), so the $n_s \times 1$ vector (with $n_s = 40$, as mentioned above) $\boldsymbol{f}^5$ has nine non-zero entries. For example, $f_1^5 = 2/12$ (because 0 occurs twice out of $n_k = 12$ total punctuation marks) and $f_4^5 = 0$ (because 3 never occurs).

The features $\boldsymbol{f}^{4,k}$ and $\boldsymbol{f}^{5,k}$ induce discrete probability distributions on the number of words in sentences and the number of words between successive punctuation marks, respectively. The expectation of the feature $\boldsymbol{f}^{5,k}$ quantifies the 'rate of punctuation' and is equal to the total number of words, relative to the total number of punctuation marks:

$$\mathbb{E}\left[\boldsymbol{f}^{5,k}\right] = \sum_{i=1}^{n_s} i f_i^{5,k} = \frac{1}{n_k} \sum_{i=1}^{n_s} i \left|\{w_l^k \in D_k^w \mid w_l^k = i\}\right| = \frac{|D_k^w|}{n_k}. \tag{2.6}$$

The feature $\boldsymbol{f}^{5,k}$ tracks word-count frequency between successive punctuation marks, without distinguishing between different punctuation marks.

To obtain $\boldsymbol{f}^{6,k}$, we compute the mean number of words between successive occurrences of the elements in ordered pairs of punctuation marks using the matrix $\boldsymbol{W}^k \in [0, n_s]^{|\Theta| \times |\Theta|}$ with entries

[8]We use the caps n_S and n_s to ensure that the features $\boldsymbol{f}^{4,k}$ and $\boldsymbol{f}^{5,k}$, respectively, have the same size (i.e., number of elements) across all documents and to mitigate the influence of outliers.

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

$$W_{ij}^k = \left\langle \{ w_l^k \in D_k^w \mid \theta_l = \theta_i \text{ and } \theta_{l+1} = \theta_j \} \right\rangle, \tag{2.7}$$

where $\langle \cdot \rangle$ denotes the sample mean of a set. The matrix for the Le Guin excerpt is

$$\boldsymbol{W} = \left[\begin{array}{cccccccccc} \textbf{!} & \textbf{"} & \textbf{(} & \textbf{)} & \textbf{,} & \textbf{.} & \textbf{:} & \textbf{;} & \textbf{?} & \textbf{...} \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 5.5 & 0 & 9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 & 0 & 0 & 7 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

We obtain $\boldsymbol{f}^{6,k}$ by flattening the matrix $\boldsymbol{W}^k$ by concatenating its rows.

There are many possible variations of the features $\boldsymbol{f}^4$, $\boldsymbol{f}^5$ and $\boldsymbol{f}^6$. For example, one need not require that punctuation-mark occurrences are successive, and one can subsequently compute the number of words or even the number of punctuation marks between the elements of an ordered pair of punctuation marks.

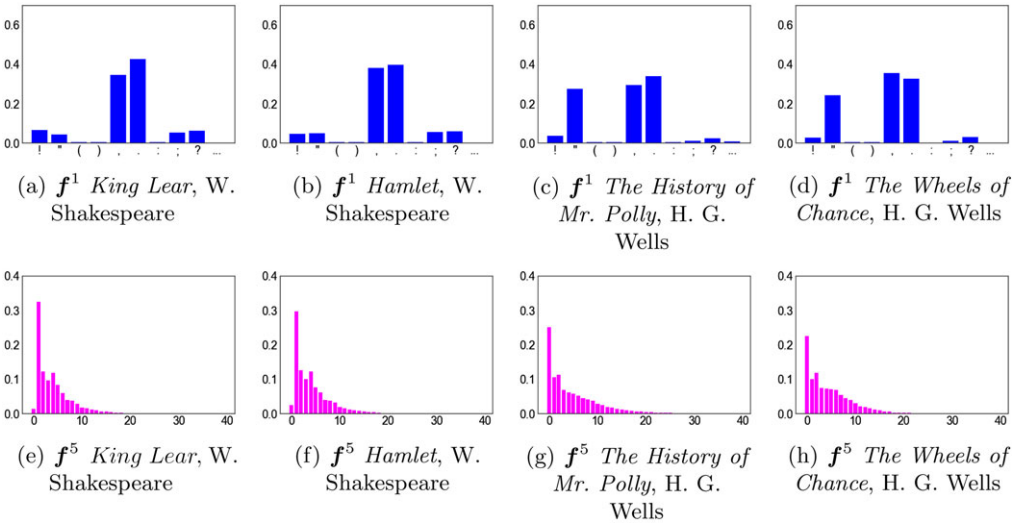
In the rest of our paper, we focus on the six features $\boldsymbol{f}^1, \ldots, \boldsymbol{f}^6$. We show example histograms of $\boldsymbol{f}^1$ (punctuation frequencies) and $\boldsymbol{f}^5$ (the frequencies of the numbers of words between successive punctuation marks) for some documents by two authors in Figure 4.

2.3 Kullback–Leibler divergence

To quantify the similarity between two discrete distributions (e.g., between the features $\boldsymbol{f}^1, \boldsymbol{f}^3, \boldsymbol{f}^4$ and $\boldsymbol{f}^5$ from different documents), we use Kullback–Leibler (KL) divergence [26], an information-theoretic measure that is related to Shannon entropy [29, 43] and ideas from maximum-likelihood theory [11, 44]. KL divergence and variants of it have been used in prior research on author recognition [2, 35, 52]. One can also consider other similarity measures, such as chi-squared distance [35] and Jensen–Shannon divergence [1, 16, 31].

Consider a discrete, finite sample space $\mathcal{X}$; and let $\boldsymbol{p} \in [0,1]^{|\mathcal{X}| \times 1}$ and $\boldsymbol{q} \in [0,1]^{|\mathcal{X}| \times 1}$ be two probability distributions on $\mathcal{X}$ that we assume are absolutely continuous with respect to each other. (In other words, we assume that an event has a non-zero probability with respect to the distribution $\boldsymbol{p}$ if and only if it has a non-zero probability with respect to $\boldsymbol{q}$.) Broadly speaking, KL divergence quantifies how close a probability distribution $\boldsymbol{p} = \{p_i\}$ is to a candidate distribution $\boldsymbol{q} = \{q_i\}$, where p_i (respectively, q_i) denotes the probability that event i occurs under $\boldsymbol{p}$ (respectively, under $\boldsymbol{q}$) [10]. The KL divergence between the probability distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ is defined as

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press



(a) $\boldsymbol{f}^1$ *King Lear*, W. Shakespeare (b) $\boldsymbol{f}^1$ *Hamlet*, W. Shakespeare (c) $\boldsymbol{f}^1$ *The History of Mr. Polly*, H. G. Wells (d) $\boldsymbol{f}^1$ *The Wheels of Chance*, H. G. Wells

(e) $\boldsymbol{f}^5$ *King Lear*, W. Shakespeare (f) $\boldsymbol{f}^5$ *Hamlet*, W. Shakespeare (g) $\boldsymbol{f}^5$ *The History of Mr. Polly*, H. G. Wells (h) $\boldsymbol{f}^5$ *The Wheels of Chance*, H. G. Wells

FIGURE 4. (a,b,c,d) Histograms of punctuation-mark frequency ($\boldsymbol{f}^1$) and (e,f,g,h) the number of words that occur between successive punctuation marks ($\boldsymbol{f}^5$) for two documents by William Shakespeare and two documents by Herbert George (H. G.) Wells.

$$d_{\mathrm{KL}}(\boldsymbol{p},\, \boldsymbol{q}) = \sum_{p_i \neq 0} p_i \log\left(\frac{p_i}{q_i}\right) \tag{2.8}$$

and satisfies three important properties:

1. $d_{\mathrm{KL}}(\boldsymbol{p},\, \boldsymbol{q}) \geqslant 0$;
2. $d_{\mathrm{KL}}(\boldsymbol{p},\, \boldsymbol{q}) = 0$ if and only if $p_i = q_i$ for all i;
3. $d_{\mathrm{KL}}(\cdot, \cdot)$ is asymmetric in its arguments.

The function 'log' denotes the natural logarithm.

To adjust for cases in which $\boldsymbol{p}$ and $\boldsymbol{q}$ are not absolutely continuous with respect to each other (e.g., one document has one or more ellipses, but another does not, resulting in unequal supports), we remove any frequency component that corresponds to a punctuation mark that is not in the common support and then distribute the weight of the removed component uniformly across the remaining frequencies. For example, suppose that $p_1 \neq 0$ but $q_1 = 0$. We then define $\widetilde{\boldsymbol{p}}$ and $\widetilde{\boldsymbol{q}}$ such that $\widetilde{\boldsymbol{p}} = \{p_i/(1-p_1) \mid i \neq 1\}$ and $\widetilde{\boldsymbol{q}} = \{q_i \mid i \neq 1\}$, and we compute $d_{\mathrm{KL}}(\widetilde{\boldsymbol{p}}, \widetilde{\boldsymbol{q}})$.

2.4 Classification models

We describe the two classification approaches that we use for author recognition (see Section 3.2) and genre recognition (see Section 4.2). Much of the existing classification work on author recognition uses machine-learning classifiers (e.g., support vector machines or neural networks) or similarity-based classification techniques (e.g., using KL divergence) [35, 45].

We use neural networks and similarity-based classification with KL divergence for both author and genre classification. Following standard practice, we split the documents in our data set into a training set and a testing set. Broadly speaking, a training set calibrates a classification model (e.g., to 'feed' a neural network and adjust its parameters), and one then uses a testing

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

set to evaluate the accuracy of a calibrated model. We ensure that all authors or genres (i.e., all 'classes') that appear in the testing set also appear in the training set; this is known as 'closed-set attribution' and is common practice in author recognition [35,45]. For a given data set, we place 80% of the documents in the training set and the remaining 20% of the documents in the testing set. (A training:testing ratio of 80:20 is a common choice.) A given data set is sometimes the entire corpus (i.e., 14,947 documents and 651 authors), and it is sometimes a subset of it. In our summary tables (see Section 3.2 and Section 4.2), we explicitly specify the sizes of the training and testing sets of our experiments.

2.4.1 *Similarity-based classification*

We label our r classes by $c_1, c_2, \ldots, c_r$ (recall that these can correspond to authors or to genres), and we denote the set of training documents for class c_j by $\mathcal{D}_j$. For each class c_j, we define a class-level feature $\boldsymbol{f}^{l,c_j}$, with $l \in \{1, \ldots, 6\}$ and $j \in \{1, \ldots, r\}$, by averaging the features across the training documents in that class. That is, the ith entry of $\boldsymbol{f}^{l,c_j}$ is

$$f_i^{l,c_j} = \frac{1}{|\mathcal{D}_j|} \sum_{k \in \mathcal{D}_j} f_i^{l,k}, \tag{2.9}$$

where $l \in \{1, \ldots, 6\}$ and we use the features $\boldsymbol{f}^{1,k}, \boldsymbol{f}^{2,k}, \ldots, \boldsymbol{f}^{6,k}$ from Section 2.2. This yields a set $\phi^k = \{\boldsymbol{f}^{1,k}, \ldots, \boldsymbol{f}^{6,k}\}$ of features for each document and a set $\phi^{c_j} = \{\boldsymbol{f}^{1,c_j}, \ldots, \boldsymbol{f}^{6,c_j}\}$ of features for each class.

To determine which class is 'most similar' to a document k in our testing set, we solve the following minimisation problem:

$$\operatorname{argmin}_{j \in \{1,\ldots,r\}} d(\phi^k, \phi^{c_j}), \tag{2.10}$$

for some choice of similarity measure $d(\cdot, \cdot)$. In our numerical experiments of Section 3, we use the KL-divergence similarity measure d_{KL} to define $d(\cdot, \cdot)$ as

$$d(\phi^k, \phi^{c_j}) = \operatorname{argmin}_{l \in \mathcal{L}} d_{\mathrm{KL}}(\boldsymbol{f}^{l,c_j}, \boldsymbol{f}^{l,k}), \tag{2.11}$$

where we restrict the set of features to those that induce probability distributions and consider each feature individually (i.e., $\mathcal{L} = \{1\}$, $\mathcal{L} = \{3\}$, $\mathcal{L} = \{4\}$ or $\mathcal{L} = \{5\}$). Recall from Section 2.2 that only the features $\boldsymbol{f}^1$, $\boldsymbol{f}^3$, $\boldsymbol{f}^4$ and $\boldsymbol{f}^5$ induce probability distributions (and thus can be compared by calculating d_{KL}). The features $\boldsymbol{f}^2$ and $\boldsymbol{f}^6$ do not induce probability distributions, so we do not consider them as inputs when we use the KL-divergence-based classifier.

2.4.2 *Neural networks*

We use feedforward neural networks with the standard backpropagation algorithm as a machine-learning classifier [25]. A neural network uses the features of a training set to automatically infer rules for recognising the classes of a testing set by adjusting the weights of each 'neuron' using a stochastic gradient-descent-based learning algorithm.

In contrast with neural networks for classical NLP classification, where it is standard to use word embeddings and employ convolutional or recurrent neural networks [27] to ensure that input vectors have equal lengths, we have already defined our features such that they have equal

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

lengths. It thus suffices for us to use feedforward neural networks. The input vector that corresponds to each document is a concatenation of the six features (or a subset thereof) that we described in Section 2.2, and the output is a vector with nonnegative entries that sum to 1, so we can interpret each of its elements as the probability that a given document belongs to a given class. We assign each document in our testing set to the class with the largest probability.

2.5 Model evaluation

For each test of a classification model, we consider a data set with a fixed number of classes (e.g., 651 classes if we perform author recognition on all authors in our corpus), a uniformly randomly sampled training set (80% of the data set), and a testing set (the remaining 20% of the data set). We measure 'accuracy' as the number of correctly assigned documents divided by the total number of documents in a testing set. For each test of a classification model, we report two quantities: (1) the accuracy of the classification model on the testing set and (2) the accuracy of a baseline classifier on the testing set. We calculate the latter by assigning each document in the testing set to a class with a probability that is proportional to that class's size in the training set.

3 Case study: Author analysis

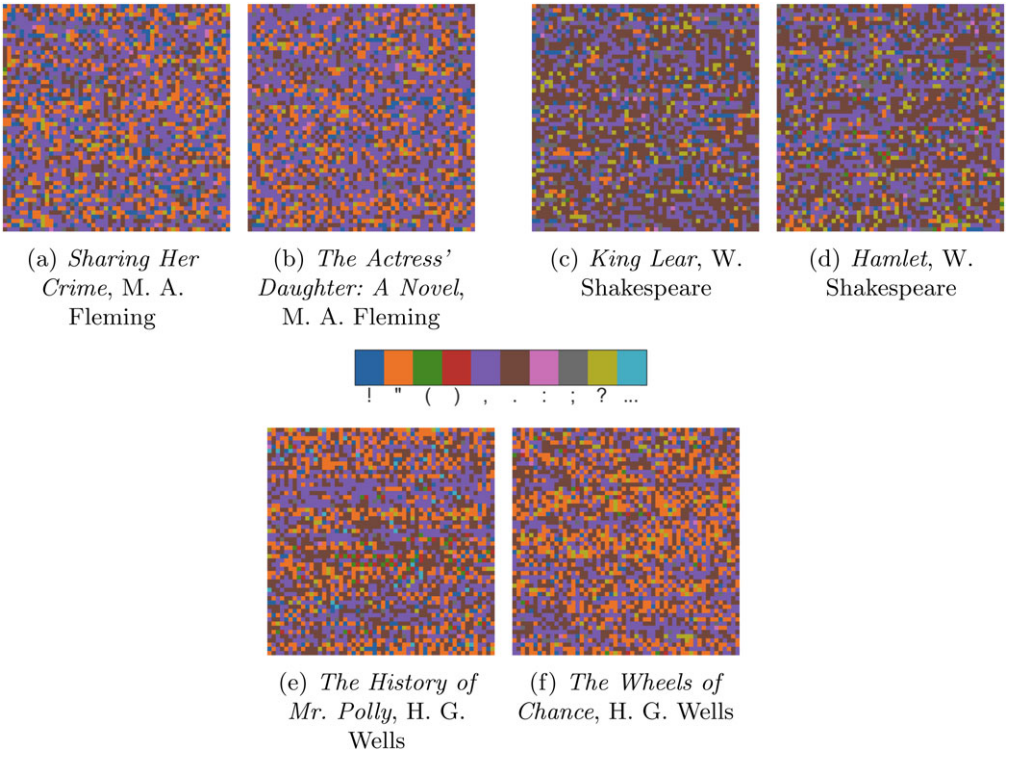
"`It is almost always a greater pleasure to come across a semicolon than a period. The period tells you that that is that; if you didn't get all the meaning you wanted or expected, anyway you got all the writer intended to parcel out and now you have to move along. But with a semicolon there you get a pleasant little feeling of expectancy; there is more to come; to read on; it will get clearer.`"

— Thomas Lewis, *Notes on Punctuation*, 1979

3.1 Consistency

We explore punctuation sequences of a few authors to gain some insight into whether certain authors have more distinguishable punctuation styles than others. (Once again, recall our cautionary note that we do not distinguish between the roles of authors and editors.) In Figure 5, we show (augmenting Figure 1) raw sequences of punctuation marks for two books by each of the following three authors: May Agnes Fleming, William Shakespeare and Herbert George (H. G.) Wells. We observe for this document sample that, visually, one can correctly guess which documents were written by the same author based only on the sequences of punctuation marks. This striking possibility was illustrated previously in A. J. Calhoun's blog entry [4], which motivated our research. From Figure 5, we see that H. G. Wells appears to use noticeably more quotation marks than the other two authors. We also observe that William Shakespeare appears to have used more periods than H. G. Wells. These observations are consistent with the histograms in Figure 4 (where we also observe that William Shakespeare appears to have used more exclamation marks and question marks than H. G. Wells), which we compute from the entire documents, so our observations from the samples in Figure 5 appear to hold throughout those documents.

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press



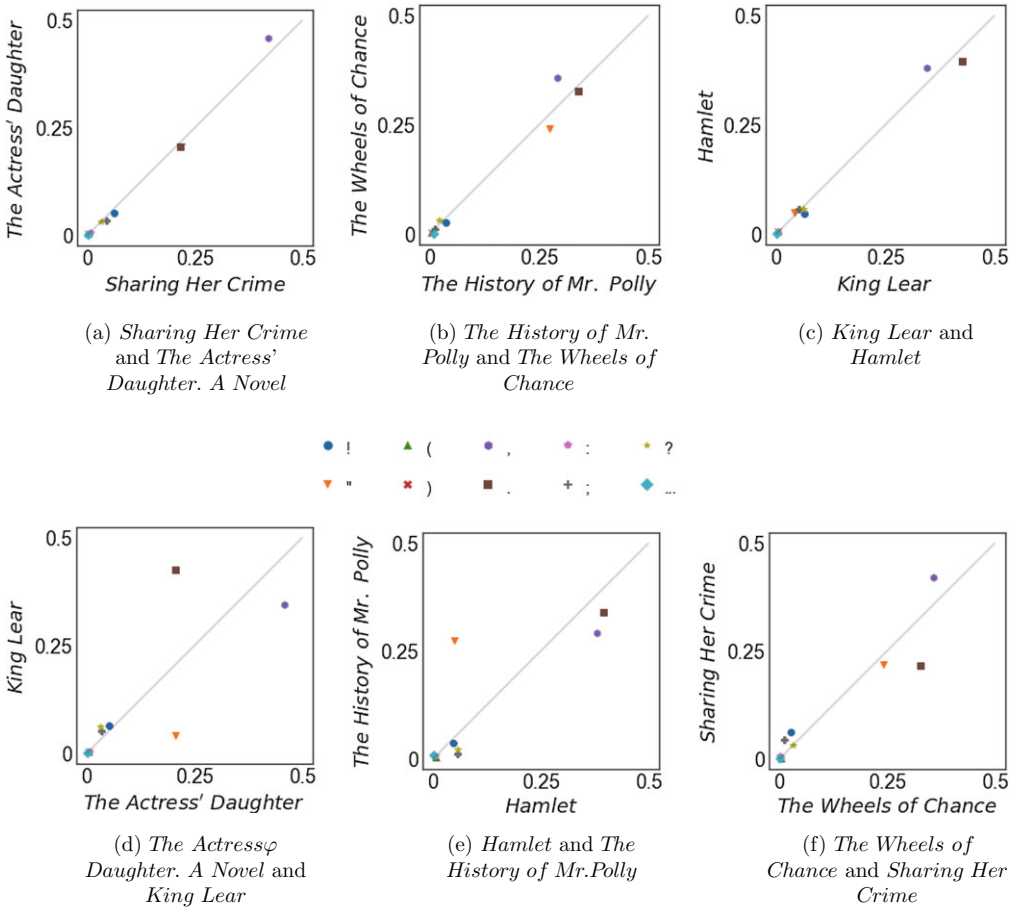
(a) *Sharing Her Crime*, M. A. Fleming
(b) *The Actress' Daughter: A Novel*, M. A. Fleming
(c) *King Lear*, W. Shakespeare
(d) *Hamlet*, W. Shakespeare
(e) *The History of Mr. Polly*, H. G. Wells
(f) *The Wheels of Chance*, H. G. Wells

FIGURE 5. Sequences of successive punctuation marks that we extract from documents by (a,b) May Agnes Fleming, (c,d) William Shakespeare and (e,f) Herbert George (H. G.) Wells. We map each punctuation mark to a distinct colour. We cap the length of each punctuation sequence at 3000 entries, which start at the midpoint of the punctuation sequence of the corresponding document.

In Figure 6, we plot examples of the punctuation frequency (i.e., $\boldsymbol{f}^1$) of one document versus that of another document by the same author (top row) and a document by a different author (bottom row). We base these plots on the 'rank order' plots of Yang et al. [51], who used such plots to illustrate the top-ranking words in various texts. In our plots, any punctuation mark (which we represent by a coloured marker) that has the same frequency in both documents lies on the grey diagonal line. Any marker above (respectively, below) the grey line is used more (respectively, less) frequently by the author on the vertical axis (respectively, horizontal axis). In these examples, we see for documents by the same author that the markers tend to be closer to the grey line than they are for documents by different authors. In Figure 6(d), for example, we observe that May Agnes Fleming used more quotation marks and commas in *The Actress' Daughter: A Novel* than William Shakespeare did in *King Lear*, whereas Shakespeare used more periods in *King Lear* than Fleming did in *The Actress' Daughter: A Novel*. One can make similar observations about panels (e) and (f) of Figure 6. These observations are consistent with those of Figures 4 and 5.

Our illustrations in Figures 5 and 6 use a very small number of documents by only a few authors. To quantify the 'consistency' of an author across all documents by that author in our corpus, we use KL divergence. In Figure 7, we show heat maps of the KL divergence between the discrete probability distributions that are induced by the features $\boldsymbol{f}^1$, $\boldsymbol{f}^3$, $\boldsymbol{f}^4$ and

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press



(a) *Sharing Her Crime* and *The Actress' Daughter. A Novel*

(b) *The History of Mr. Polly* and *The Wheels of Chance*

(c) *King Lear* and *Hamlet*

(d) *The Actressφ Daughter. A Novel* and *King Lear*

(e) *Hamlet* and *The History of Mr.Polly*

(f) *The Wheels of Chance* and *Sharing Her Crime*

FIGURE 6. Scatter plots of frequency vectors (i.e., $\boldsymbol{f}^1$) of punctuation marks to compare books from the same author: (a) *Sharing Her Crime* and *The Actress' Daughter: A Novel* by May Agnes Fleming, (c) *King Lear* and *Hamlet* by William Shakespeare and (e) *The History of Mr. Polly* and *The Wheels of Chance* by H. G. Wells. Scatter plots of frequency vectors of punctuation marks to compare books from different authors: (b) *The Actress' Daughter: A Novel* and *King Lear*, (d) *Hamlet* and *The History of Mr. Polly* and (f) *The Wheels of Chance* and *Sharing Her Crime*. We represent each punctuation mark by a coloured marker, with coordinates from the punctuation frequencies in a vector that is associated with each document. The grey line represents the identity function. More similar frequency vectors correspond to dots that are closer to the grey line.

$\boldsymbol{f}^5$. Recall from Section 2.2 that $\boldsymbol{f}^1$ quantifies the frequency of punctuation marks, $\boldsymbol{f}^3$ quantifies the frequency of successive occurrences of each pair of punctuation marks, $\boldsymbol{f}^4$ quantifies the frequency of sentence lengths and $\boldsymbol{f}^5$ quantifies the word-count frequency between successive punctuation marks without distinguishing between different punctuation marks. We compute each feature for each document in our corpus. We define the 'consistency' of an author relative to a feature as the mean KL divergence for that feature across all pairs of documents by that author. That is, the consistency of author a with respect to feature $\boldsymbol{f}^i$ is

$$\mathcal{C}_{\boldsymbol{f}^i}(a) = \frac{2}{|\mathcal{D}_a - 1||\mathcal{D}_a|} \sum_{k,k' \in \mathcal{D}_a} d_{\mathrm{KL}}(\boldsymbol{f}^{i,k}, \boldsymbol{f}^{i,k'}), \tag{3.1}$$

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

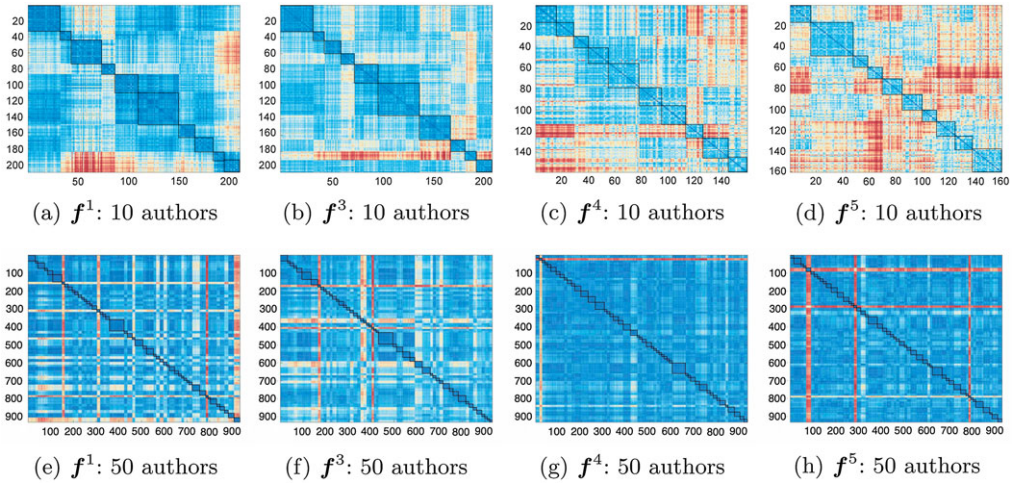


FIGURE 7. Heat maps of the KL divergence between different authors for the features (a,e) $\boldsymbol{f}^1$, (b,f) $\boldsymbol{f}^3$, (c,g) $\boldsymbol{f}^4$ and (d,h) $\boldsymbol{f}^5$. We show the 10 most-consistent authors (see the main text for our notion of 'consistency') for each feature in the top row and the 50 most-consistent authors for each feature in the bottom row. The diagonal blocks that we enclose in black indicate documents by the same author. Authors can differ across panels, because author consistency can differ across features. The colours scale (nonlinearly) from dark blue (corresponding to a KL divergence of 0) to dark red (corresponding to the maximum value of the KL divergence in the depicted matrix). For ease of exposition, we suppress colour bars (they span the interval $[0, 3.35]$), as we seek to illustrate the presence and/or absence of high-level structure. When determining the 10 most-consistent authors, we exclude the authors 'United States. Warren Commission' (see row 7 in Table A.1) in panels (a)–(c) and we exclude the author 'United States. Central Intelligence Agency' (see row 198 in Table A.1) in panel (d). In each case, we replace the excluded author with the next-most-consistent author and proceed from there. Works by these two authors consist primarily of court testimonies or lists of definitions and facts (with rigid styles); they manifested as pronounced dark-red stripes that masked salient block structures.

where $\mathcal{D}_a$ denotes the set of documents by author a. For each feature in Figure 7, we show the 10 (respectively, 50) most-consistent authors in the top row (respectively, bottom row). Diagonal blocks with black outlines correspond to documents by the same author. Although there appears to be greater similarity within diagonal blocks than between them for several of the authors, it is difficult to interpret the heat maps when there are many authors (and it becomes increasingly difficult as one considers progressively more authors).

In Figure 8, we show author consistency in our entire corpus for the features $\boldsymbol{f}^1$, $\boldsymbol{f}^3$, $\boldsymbol{f}^4$ and $\boldsymbol{f}^5$. In each panel, we show a baseline (in blue), which we obtain by choosing, uniformly at random, 1000 ordered pairs of documents by distinct authors and computing the mean KL divergences between these document pairs for these features. One pair is a single element of an off-diagonal block of a matrix like those in Figure 7.

We order each panel from the least-consistent author to the most-consistent author. Authors can differ across panels, because the consistency measure (3.1) is a feature-dependent quantity. We observe in all panels of Figure 8 that most authors are more consistent on average than the baseline. (Visually, the black curve lies below the blue horizontal line for most authors.) The differences between authors relative to the baseline are most pronounced for the feature $\boldsymbol{f}^3$

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

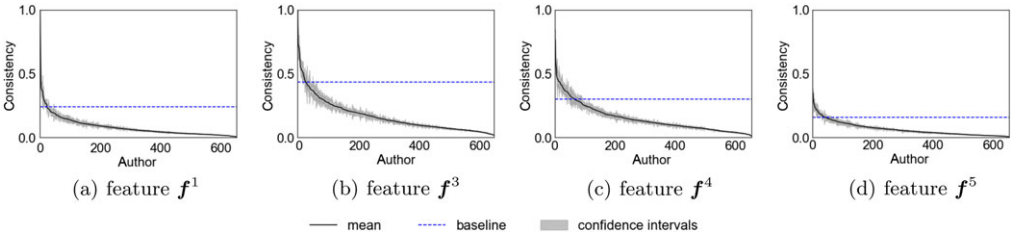


FIGURE 8. Evaluations of author consistency. In each panel, we show author consistency (3.1) for the features (a) f^1, (b) f^3, (c) f^4 and (d) f^5 using a solid black curve. In grey, we plot confidence intervals of the KL divergence across ordered pairs of documents for each author. To compute the confidence intervals, we assume for each author that the KL-divergence values across pairs of distinct documents are normally distributed. There are at least 10 documents by each author in our corpus (see Section 2.1), so the number of KL-divergence values across pairs of distinct documents by a given author is at least 90. The dotted blue line indicates a consistency baseline, which we obtain by choosing, uniformly at random, 1000 ordered pairs of documents by distinct authors and computing the mean KL divergences between these document pairs.

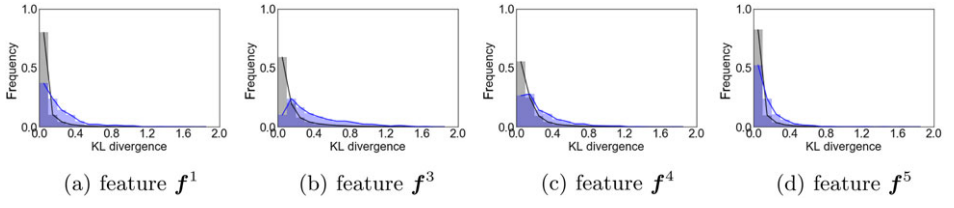


FIGURE 9. Distributions of KL divergence for authors. In each panel, we show the distributions of KL divergence between all ordered pairs of documents in the corpus by the same author (in black) and between 1000 ordered pairs of documents by distinct authors (in blue). We choose the ordered pairs uniformly at random from the set of all ordered pairs of documents by distinct authors. Each panel corresponds to a distinct feature. The means of the distributions of each panel are (a) 0.0828 (black) and 0.240 (blue), (b) 0.167 (black) and 0.433 (blue), (c) 0.149 (black) and 0.275 (blue) and (d) 0.0682 (black) and 0.154 (blue).

(see Table 1). This suggests that f^3 may carry more information than our other five features about an author's idiosyncratic style. We return to this observation in Section 3.2.

In Figure 9, we show the distributions of the KL-divergence values between documents by the same authors (in black) and between documents by distinct authors (in blue). In Figure 8, we used the former to compute author consistency (by taking the mean of the values for each author) and the latter to compute the consistency baseline (by taking the mean of all values). For all features, we see from a Kolmogorov–Smirnov (KS) test that the difference between the empirical distributions is statistically significant. (In all cases, the p-value is less than or equal to 1.218×10^{-79}.)

3.2 Author recognition

We use the classification techniques from Section 2.4 to perform author recognition. We show our results using KL divergence (see Section 2.4.1) in Table 2 and using neural networks

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

Table 2. *Results of our author-recognition experiments using a classification based on KL divergence (see Section 2.4.1) for author samples of various sizes and with the individual features* $\boldsymbol{f}^1$, $\boldsymbol{f}^3$, $\boldsymbol{f}^4$ *and* $\boldsymbol{f}^5$ *as input. We measure accuracy as the number of correctly assigned documents divided by the total number of documents in the testing set. See Section 2.5 for a description of the baseline.*

			Accuracy for the testing set				
No. authors	Training size	Testing size	$\boldsymbol{f}^1$	$\boldsymbol{f}^3$	$\boldsymbol{f}^4$	$\boldsymbol{f}^5$	baseline
10	216	55	0.69	0.74	0.52	0.63	0.21
50	834	209	0.54	0.66	0.30	0.31	0.029
100	2006	502	0.37	0.49	0.25	0.23	0.019
200	3549	888	0.30	0.47	0.16	0.20	0.0079
400	7439	1860	0.27	0.41	0.15	0.16	0.0047

(see Section 2.4.2) in Table 3. In each table, we specify the number of authors ('No. authors'), the number of documents in the training set ('Training size'), the number of documents in the testing set ('Testing size'), the accuracy of the test using various sets of features and the baseline accuracy (as defined in Section 2.5). Each row in a table corresponds to an experiment on a set of distinct authors, which we choose uniformly at random. (The set consists of the entire corpus when the number of authors is 651.) For a given number of authors, we use the same sample across both tables to allow a fair comparison.

In Table 2, we show our classification results using KL divergence for each individual feature (i.e., $\boldsymbol{f}^1$, $\boldsymbol{f}^3$, $\boldsymbol{f}^4$ and $\boldsymbol{f}^5$) that induces a probability distribution. As we consider more authors, the accuracy for the testing set tends to decrease significantly. The issue of developing a method that performs well as one increases the number of authors is an open problem in author recognition even when using words from text [35], and we are exploring stylistic signatures from punctuation only, a much smaller set of information. Remarkably, we are able to achieve an accuracy of about 66% on a sample of 50 authors using only the feature $\boldsymbol{f}^3$. This is consistent with the plots in Figure 8, where $\boldsymbol{f}^3$ gave the best improvement from the baseline.

In Table 3, we show our classification results using a one-layer neural network with 2000 neurons for various sets of inputs (which, in contrast to when one uses KL divergence, do not have to be features that induce probability distributions). We also observe in Table 3 that the accuracy for the testing set tends to decrease significantly as we increase the number of authors. Overall, however, the neural network outperforms our KL-divergence-based classification. We achieve an accuracy of about 62% when using only $\boldsymbol{f}^3$ and an accuracy of about 72% when using all features on a sample of 651 authors (i.e., on the entire corpus). Interestingly, in some of our experiments, using the feature set $\{\boldsymbol{f}^1, \boldsymbol{f}^3, \boldsymbol{f}^4, \boldsymbol{f}^5\}$ gives slightly better accuracy than using all features.

Based on several repetitions of our experiments, the accuracy results in Tables 2 and 3 seem to be robust with respect to (1) different author samples of the same size and (2) different training and testing samples for a given author sample. However, the heterogeneity in accuracy across

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

Table 3. *Results of our author-recognition experiments using a one-layer, 2000-neuron neural network (see Section 2.4.2) for author samples of various sizes and with different features or sets of features as input:* $\boldsymbol{f}^1$, $\boldsymbol{f}^3$, $\boldsymbol{f}^4$, $\boldsymbol{f}^5$, $\{\boldsymbol{f}^1, \boldsymbol{f}^3, \boldsymbol{f}^4, \boldsymbol{f}^5\}$ *and* $\{\boldsymbol{f}^1, \boldsymbol{f}^2, \boldsymbol{f}^3, \boldsymbol{f}^4, \boldsymbol{f}^5, \boldsymbol{f}^6\}$ *(which we label as 'All'). We measure accuracy as the number of correctly assigned documents divided by the total number of documents in the testing set. See Section 2.5 for a description of the baseline.*

No. authors	Training size	Testing size	Accuracy for the testing set						
			$\boldsymbol{f}^1$	$\boldsymbol{f}^3$	$\boldsymbol{f}^4$	$\boldsymbol{f}^5$	$\{\boldsymbol{f}^1, \boldsymbol{f}^3, \boldsymbol{f}^4, \boldsymbol{f}^5\}$	All	Baseline
10	216	55	0.89	0.93	0.64	0.80	0.89	0.87	0.21
50	834	209	0.65	0.81	0.44	0.49	0.81	0.82	0.029
100	2006	502	0.55	0.79	0.37	0.39	0.79	0.80	0.019
200	3549	888	0.46	0.71	0.23	0.32	0.71	0.75	0.0079
400	7439	1860	0.39	0.70	0.23	0.27	0.71	0.73	0.0047
600	11102	2776	0.37	0.70	0.21	0.25	0.61	0.74	0.0029
651	11957	2990	0.36	0.62	0.20	0.23	0.67	0.72	0.0024

different author samples with the same number of authors is more pronounced than the heterogeneity in accuracy that we observe from different training and testing samples for a given author sample, as different author samples can yield different numbers of documents (see Figure 2). Such heterogeneity across different author samples decreases as one increases the number of authors.

To the best of our knowledge, most attempts thus far at author recognition of literary documents have used data sets that are significantly smaller than our corpus [15, 35]. One recent example of author analysis from a corpus from Project Gutenberg is the one in Qian et al. [40]. Their corpus consists of 50 authors (with their choices of authors based on a popularity criterion) and 900 single-paragraph excerpts for each author. (For a given author, they extracted their excerpts from several books.) Using word-based features and machine-learning classifiers, they achieved an accuracy of about 89.2% using 90% of their data for training and 10% of it for testing.

4 Case study: Genre analysis

"`Cut out all those exclamation marks. An exclamation mark is like laughing at your own jokes.`"

— Attributed to F. Scott Fitzgerald, as conveyed by Sheilah Graham and Gerold Frank in *Beloved Infidel: The Education of a Woman*, 1958

"'`Multiple exclamation marks,' he went on, shaking his head, 'are a sure sign of a diseased mind.`'"

— Terry Pratchett, *Eric*, 1990

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

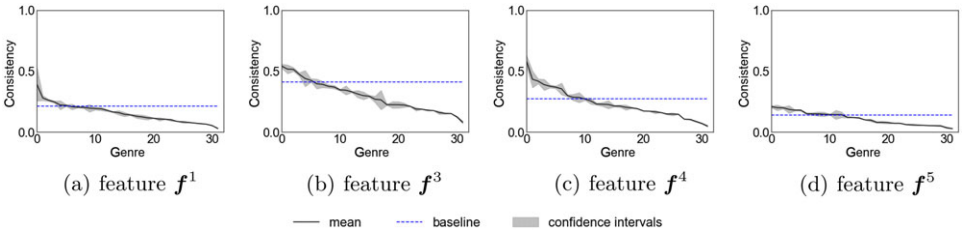


FIGURE 10. Evaluation of genre consistency. In each panel, we show the genre consistency (specifically, we use equation (3.1), but with genres, instead of authors) for (a) f^1, (b) f^3, (c) f^4 and (d) f^5 as a solid black curve. In grey, we show confidence intervals of the KL divergence across ordered pairs of documents for each genre. To compute the confidence intervals, we assume for each genre that the KL-divergence values across pairs of distinct documents are normally distributed. There are at least 10 documents for each genre in our corpus (see the introduction of Section 4), so the number of KL-divergence values across pairs of distinct documents for each genre is at least 90. The dotted blue line indicates a consistency baseline, which we obtain by choosing, uniformly at random, 1000 ordered pairs of documents from distinct genres and computing the mean KL divergences between these document pairs.

We now use genres as our classes. Among the 121 genre ('bookshelf') labels that are available in Project Gutenberg,[9] we keep those that include at least 10 documents. Among the remaining genres, we manually select 32 relatively unspecialised genre labels. We show this final list of genres in Appendix A. This yields a data set with 2413 documents.

4.1 Consistency

In Figure 10, we show consistency plots (of the same type as in Figure 8), but now we use genres (instead of authors) as our classes. We observe that the KL-divergence consistency relative to the baseline is less pronounced for genres than it was for authors. Nevertheless, most genres are more consistent than the baseline, and f^3 appears to be the most helpful of our features for evaluating a genre's punctuation style.

In Figure 11, we show the distributions of KL divergence between documents from the same genre (in black) and between documents from different genres (in blue). One can use the former to compute genre consistency in Figure 10 (by taking the mean of the values for each genre) and the latter to compute the consistency baseline in Figure 10 (by taking the mean of all values). For all features, we see from a KS test that the difference between the empirical distributions is statistically significant. (In all cases, the p-value is less than or equal to 2.247×10^{-36}.)

4.2 Genre recognition

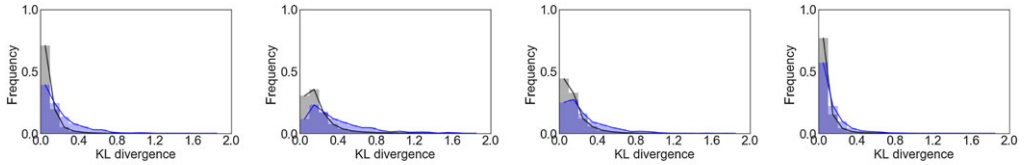
We perform genre recognition using neural networks and show our results in Table 4. We are less successful at genre detection than we were at author detection. This is consistent with our genre consistency plots (see Figure 10), which indicated a smaller differentiation from the baseline than in our author consistency plots (see Figure 8). Our highest accuracy for genre recognition in the experiment that we show in Table 4 is 65%; we achieve it when using only the feature

[9]Every document in our corpus has at most one genre, but most documents are not assigned a genre.

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

Table 4. *Results of our genre-recognition experiments using a one-layer, 2000-neuron neural network (see Section 2.4.2) with different features or sets of features as input:* $\boldsymbol{f}^1$, $\boldsymbol{f}^3$, $\boldsymbol{f}^4$, $\boldsymbol{f}^5$, $\{\boldsymbol{f}^1, \boldsymbol{f}^3, \boldsymbol{f}^4, \boldsymbol{f}^5\}$ *and* $\{\boldsymbol{f}^1, \boldsymbol{f}^2, \boldsymbol{f}^3, \boldsymbol{f}^4, \boldsymbol{f}^5, \boldsymbol{f}^6\}$ *(which we label as 'All'). We measure accuracy as the number of correctly assigned documents divided by the total number of documents in the testing set. See Section 2.5 for a description of the baseline.*

No. genres	Training size	Testing size	Accuracy for the testing set						
			$\boldsymbol{f}^1$	$\boldsymbol{f}^3$	$\boldsymbol{f}^4$	$\boldsymbol{f}^5$	$\{\boldsymbol{f}^1, \boldsymbol{f}^3, \boldsymbol{f}^4, \boldsymbol{f}^5\}$	All	Baseline
32	1930	483	0.56	0.65	0.37	0.40	0.61	0.64	0.094



(a) feature $\boldsymbol{f}^1$ (b) feature $\boldsymbol{f}^3$ (c) feature $\boldsymbol{f}^4$ (d) feature $\boldsymbol{f}^5$

FIGURE 11. Distributions of KL divergence for genre. In each panel, we show the distributions of KL divergence between all ordered pairs of documents in our corpus from the same genre (in black) and between 1000 ordered pairs of documents from distinct genres (in blue). We choose the ordered pairs uniformly at random from the set of all ordered pairs of documents from distinct genres. Each panel corresponds to a distinct feature. The means of the distributions of each panel are (a) 0.102 (black) and 0.215 (blue), (b) 0.206 (black) and 0.412 (blue), (c) 0.154 (black) and 0.272 (blue) and (d) 0.0821 (black) and 0.138 (blue).

$\boldsymbol{f}^3$ as input. Our accuracy results are similar for different samples of the training and testing sets (although the order is sometimes different for feature sets that yield similar accuracies).

5 Case study: Temporal analysis

"'Whatever it is that you know, or that you don't know, tell me about it. We can exchange tirades. The comma is my favorite piece of punctuation and I've got all night.'"

— Rasmenia Massoud, *Human Detritus*, 2011

"Who gives a @!#?@! about an Oxford comma?
I've seen those English dramas too
They're cruel"

— Vampire Weekend, *Oxford Comma*, 2008

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

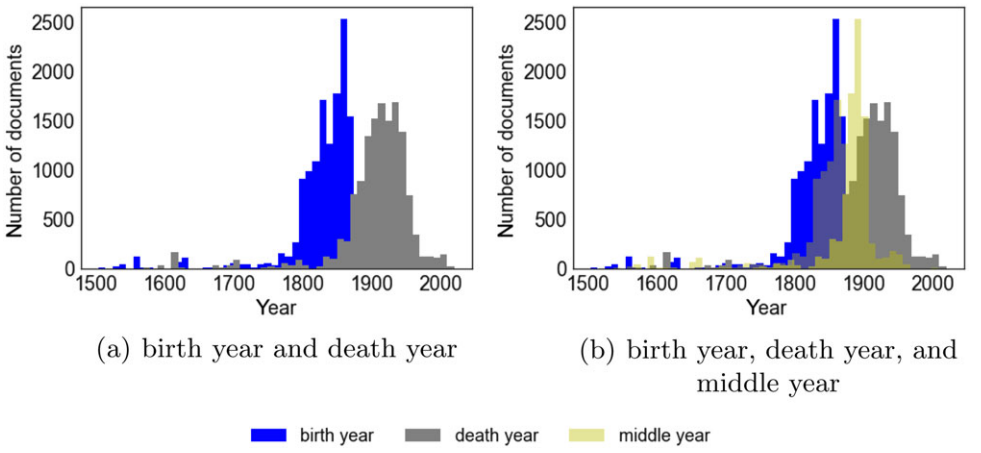


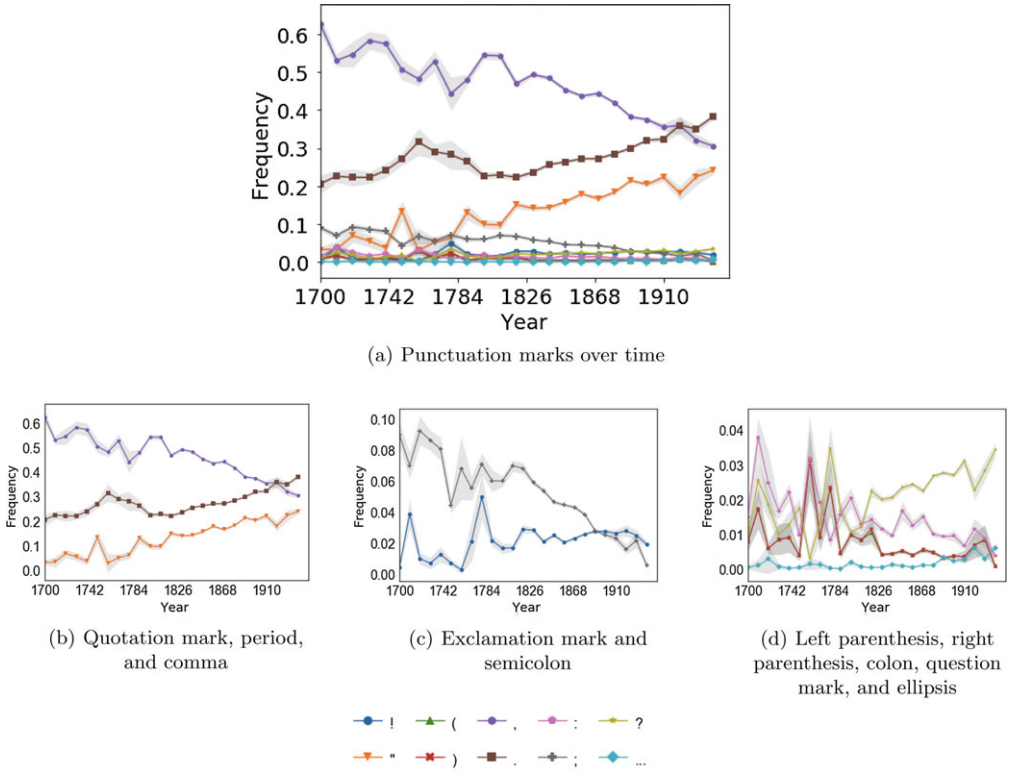
FIGURE 12. Distribution of author dates over time in our corpus. The bars represent the number of documents by author birth year (blue) and death year (grey) split into bins, where each bin represents a 10-year period. (We start at 1500.) For ease of visualisation, we only show documents for authors who were born in 1500 or later. (Only six of our authors for whom we have birth years were born before 1500.) We determine the 'middle year' of an author by taking the mean of the birth year and the death year if they are both available. If we know only the birth year, we assume that the middle year of an author is 30 years after the birth year; if we know only the death year, we assume that the middle year is 30 years prior to the death year.

We perform experiments to obtain preliminary insight into how punctuation has changed over time. In our corpus, we have access to the birth year and death year of 614 and 615 authors, respectively, of our 651 authors. We have both the birth and death years for 607 authors. In Figure 12, we show the distribution of the number of documents by author birth year, death year and 'middle year'.[10] (See the caption of Figure 12 for the definition of middle year.) We restrict our analysis to authors with a middle year between 1500 and 2012. Of the authors for whom we possess either a birth year or a death year, 616 of them have a middle year between 1500 and 2012. We show the evolution of punctuation marks over time for these 616 authors in Figure 13 and Figure 14, and we examine the punctuation usage of specific authors over time in Figure 15. Based on our experiments, it appears from Figure 13 that the use of quotation marks and periods has increased over time (at least in our corpus), but that the use of commas has decreased over time. Less noticeably, the use of semicolons has also decreased over time.[11] In Figure 14, we observe that the punctuation rate (which is given by formula (2.6)) tends to decrease over time in our corpus. However, this observation requires further statistical testing, especially given the large variance in Figure 14. Because of our relatively small number of documents per author and the uneven distribution of documents in time, our experiments in Figure 15 give only preliminary insights into the temporal evolution of punctuation, which merits a thorough analysis with a much larger (and more appropriately sampled)

[10]We use 'middle year' as a proxy for 'publication year', which is unavailable in the metadata of Project Gutenberg. Our results are qualitatively similar when we use birth year or death year (instead of middle year).

[11]See [49] for a 'biography' of the semicolon, which reportedly was invented in 1494.

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press



(a) Punctuation marks over time

(b) Quotation mark, period, and comma

(c) Exclamation mark and semicolon

(d) Left parenthesis, right parenthesis, colon, question mark, and ellipsis

FIGURE 13. Mean frequencies of punctuation marks in each middle year versus the middle years of authors. Recall that $\boldsymbol{f}^{1,k}$ is the frequency of punctuation marks for document k. We bin middle years into 10-year periods that start at 1700. In (a), we show the temporal evolution of all punctuation marks. For clarity, we also separately plot (b) the three punctuation marks with the largest frequencies in the final year of our data set, (c) the next two most-frequent punctuation marks and (d) the remaining punctuation marks. The grey shaded area indicates confidence intervals. To compute the confidence intervals, we assume for each year that the values of $\boldsymbol{f}^{1,k}$ are normally distributed.

corpus. Nevertheless, our case study illustrates the potential for studying the temporal evolution of punctuation styles of authors, genres and literature (and other text).

6 Conclusions and discussion

"La punteggiatura è come l'elettroencefalogramma di un cervello che sogna — non dà le immagini ma rivela il ritmo del flusso sottostante."

— Andrea Moro, *Il Segreto di Pietramala*, 2018

We explored whether punctuation is a sufficiently rich stylistic feature of text to distinguish between different authors and between different genres, and we also examined how punctuation has evolved over time. Using a large corpus of documents from Project Gutenberg, we observed that simple punctuation-based quantitative features (which account for both frequency and order

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

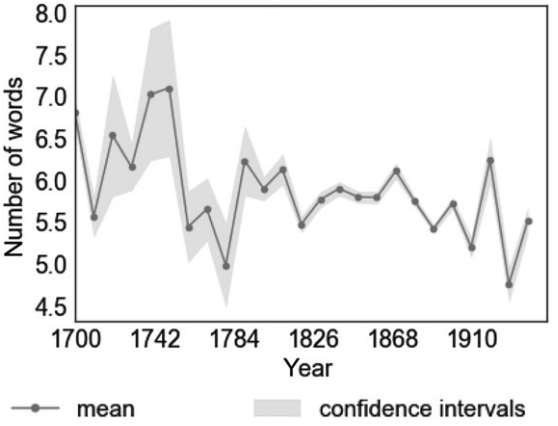
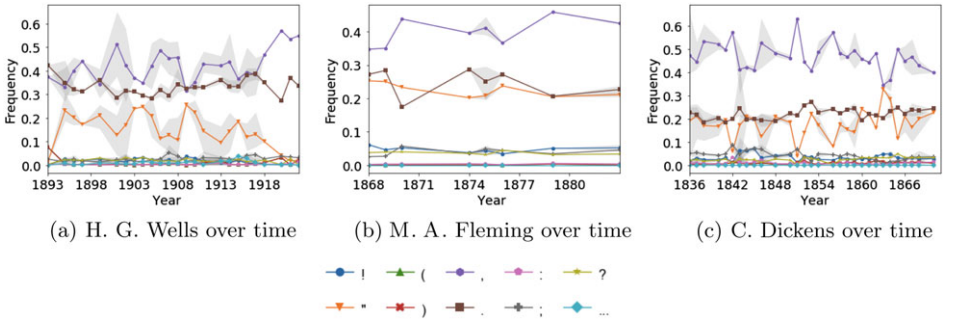


FIGURE 14. Temporal evolution of the mean number of words between two consecutive punctuation marks (i.e., $\mathbb{E}\left[\boldsymbol{f}^{5,k}\right]$ from formula (2.6)) versus author middle years, which we bin into 10-year periods that start at 1700. The grey shaded area indicates confidence intervals. To compute the confidence intervals, we assume for each year that the values of $\mathbb{E}\left[\boldsymbol{f}^{5,k}\right]$ are normally distributed. This reflects how the punctuation rate in our corpus has changed over time.



(a) H. G. Wells over time (b) M. A. Fleming over time (c) C. Dickens over time

FIGURE 15. Mean frequencies of punctuation marks in each publication year versus the publication years of works by (a) Herbert George (H. G.) Wells, (b) May Agnes Fleming and (c) Charles Dickens. Recall that $\boldsymbol{f}^{1,k}$ is the frequency vector for the punctuation marks in document k. The grey shaded area indicates the minimum and maximum value of $\boldsymbol{f}^{1,k}$ for each year. Because of the small sample sizes, we do not show confidence intervals.

of punctuation marks) can distinguish accurately between the styles of different authors. These features can also help distinguish between genres, although they do so less successfully than for authors. One feature, which we denoted by $\boldsymbol{f}^3$, measures the frequency of successive punctuation marks (and thereby partly accounts for the order in which punctuation marks occur). Among the features that we studied, it revealed the most information about punctuation style across all of our experiments. It is worth noting that, unlike the feature $\boldsymbol{f}^2$, which also accounts for the order of punctuation marks, $\boldsymbol{f}^3$ gives less weight to rare events and more weight to frequent events (see equation (2.3)). This characteristic of $\boldsymbol{f}^3$, in concert with the fact that it accounts partly for the order of punctuation marks, may explain some of its success in our experiments. It would be interesting to investigate whether particular entries of $\boldsymbol{f}^3$ have more explanatory power than others, and it is also worth exploring the accuracy of tasks like author recognition

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

as a function of the lengths of the punctuation sequences that one extracts from a document. The latter exploration may shed light into how much of a 'punctuation signal' is necessary to determine an author's stylistic footprint. In preliminary explorations, we also observed changes in punctuation style over time, but it is necessary to conduct more thorough investigations of temporal usage patterns.

To assess whether our observations extend beyond our Project Gutenberg corpus, it is necessary to conduct further experiments (e.g., on a larger corpus, across different e-book sources, and so on). For example, it is desirable to repeat our analysis using the 'Text data' level of granularity in the recently introduced '*Standardized Project Gutenberg Corpus*' [15]. Additionally, although we associate documents to authors throughout our paper as an expository shortcut, we reiterate that authors and editors both influence a document's writing and punctuation style, and we do not distinguish between the two in our analysis. It would be interesting (although daunting and computationally challenging for someone to do it with Project Gutenberg) to try to gauge whether and how much different editors affect authorial style.[12] It is also worth reiterating that Project Gutenberg has limitations with the cleanliness of its data. (See our discussion in Section 2.1 for examples of such issues.) These issues may be inherited from the e-books themselves, and they can also arise from how the documents were entered into Project Gutenberg. Although we extensively cleaned the data from Project Gutenberg to ameliorate some of its limitations, important future work is comparing documents that one extracts from Project Gutenberg with the same documents from other data sources.

Our framework allows the exploration of numerous other fascinating ideas. For example, we expect that it will be fruitful to examine higher-order Markov chains when accounting for punctuation order. Additionally, we look forward to extensions of our work that explore other features, such as the number of words between elements in ordered pairs of punctuation marks (even when they are not successive) and different ways of measuring punctuation frequency [17] and sentence length [48]. It is also worthwhile to try to quantify how large a sample of a document is necessary to correctly identify its features of punctuation style. If this size is sufficiently small, it may even be possible to identify punctuation style from collections of short text (such as tweets by politicians with limited coherence). It is also likely to be useful to exploit machine-learning classifiers that can take raw punctuation sequences (rather than features that one produces from them, as in the present work) as input and exploit 'long-range correlations' [12] between punctuation marks.

Building on our analysis, it will be interesting to investigate other aspects of stylometry — such as author pacing or the influence on an author of gender, culture, other demographics, local history or other aspects of humanity — and to compare the results of punctuation-based stylometry with existing (word-based) approaches in NLP on the same tasks. One can also explore how successful punctuation-based features are at plagiarism detection and investigate whether the punctuation in a part of a document (e.g., one chapter) is representative of the punctuation in a whole document. Further investigations of a punctuation-based approach to stylometry will also provide an opportunity to apply other methods for analysing categorical time series (e.g., an extension of rough-path signatures [8, 32] to categorical time series).

[12]Such an analysis may be easier with academic papers, as one can compare papers on arXiv to their published versions.

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

We anticipate that approaches that build on the ideas in our paper will be useful for a variety of applications, including analysis of stylistic differences in punctuation between politicians from different political parties [5] and comparisons between different editions of the same book. It will also be interesting to explore the effects of an editor's or journal's style on documents by a given author (an especially relevant study, in light of the potential to confound such contributions in corpuses like Project Gutenberg), as well as the effects of a translator's style on documents. We envisage that an application that focuses on translations is particularly well-suited to punctuation-based stylometry, as punctuation marks are supra-linguistic in nature [28] and thus depend far less than words on the specific choice of language. We also imagine a variety of potential commercial applications (e.g., using online data sources) of time-series analysis of symbols without the use of words.

Acknowledgements

The original inspiration for this project was Adam Calhoun's blog entry [4] and its striking visualisations of punctuation sequences. We thank Mariano Beguerisse Díaz, Arthur Benjamin, Bryan Bischof, Chris Brew, Cynthia Gong, Joanna Innes, Jalil Kazemitabar, Aisling Kelliher, Terry Lyons, Ursula Martin, Stephen Pulman, Massimo Stella, Adam Tsakalidis, Dmitri Vainchtein, Bo Wang and two anonymous referees for helpful comments. Several attendees at SDH's 60th birthday workshop (see https://www.maths.ox.ac.uk/groups/mathematical-finance/sam-howisons-60th-birthday-workshop-2018) also made helpful comments. For part of this project, MB was supported by The Alan Turing Institute under Engineering and Physical Sciences Research Council (EPSRC) grant EP/N510129/1. MAP and SDH thank their students and postdocs for putting up with many long discussions about punctuation when they perhaps should have been discussing other elements of their scholarship. (It was inevitable that we would eventually write an article like this.) MAP thanks SDH for his collaboration and friendship, and he wishes him a very happy birthday filled with British spelling, the word 'which' (and occasionally 'that'), and minimal commas (and parenthetical remarks).

Conflict of Interest

None.

References

[1] ALTMANN, E. G., DIAS, L. & GERLACH, M. (2017) Generalized entropies and the similarity of texts. *J. Stat. Mech. Theory Exp.* **1**, 014002.

[2] ARUN, R., SURESH, V. & MADHAVAN, C. E. V. (2009) Stopword graphs and authorship attribution in text corpora. In: *Proceedings of the 2009 IEEE International Conference on Semantic Computing*, pp. 192–196.

[3] CALHOUN, A. J. (2016) Punctuation code. Available at https://github.com/adamjcalhoun/punctuation.

[4] CALHOUN, A. J. (2016) Punctuation in novels. Available at https://medium.com/~@neuroecology/punctuation-in-novels-8f316d542ec4#.brev0b3w1.

[5] CALHOUN, A. J. (2016) What does punctuation tell us about Republicans and Democrats? Available at https://medium.com/@neuroecology/what-does-punctuation-tell-us-about-republicans-and-democrats-bd46b9f98220.

[6] CAN, F. & PATTON, J. M. (2004) Change of writing style with time. *Comput. Human.* **38**, 61–82.

[7] CHASKI, C. E. (2001) Empirical evaluation of language-based author identification techniques. *Forensic Linguist.* **8**, 1–65.
[8] CHEVYREVA, I. & KORMILITZIN, A. (2016) A primer on the signature method in machine learning. arXiv:1603.03788.
[9] CHIANG, H., GE, Y. & WU, C. (2015) *Classification of Book Genres by Cover and Title*. Class report, Computer Science 229, Stanford University. Available at http://cs229.stanford.edu/proj2015/127_report.pdf.
[10] COVER, T. M. & THOMAS, J. A. (1991) *Elements of Information Theory*, John Wiley & Sons, Inc., New York City, NY, USA.
[11] DUDA, R. O., HART, P. E. & STORK, D. G. (2001) *Pattern Classification*, John Wiley & Sons, Inc., New York City, NY, USA.
[12] EBELING, W. & PÖSCHEL, T. (1994) Entropy and long-range correlations in literary English. *Europhysics Lett.* **26**, 241–246.
[13] FORSYTH, R. S. (1999) Stylochronometry with substrings, or: a poet young and old. *Literary Linguist. Comput.* **14**, 467–477.
[14] FOWLER, H. W. & FOWLER, F. G. (1906) *The King's English*, Oxford University Press, Oxford, UK.
[15] GERLACH, M. & FONT-CLOS, F. (2020) A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. Entropy **22**, 126.
[16] GERLACH, M., FONT-CLOS, F. & ALTMANN, E. G. (2016) Similarity of symbol frequency distributions with heavy tails. *Phys. Rev. X* **6**, 021009.
[17] GRIEVE, J. (2007) Quantitative authorship attribution: an evaluation of techniques. *Literary Linguist. Comput.* **22**, 251–270.
[18] HART, M. S. (1971) Project Gutenberg. Available at https://www.gutenberg.org.
[19] HARTMAN, C. O. (2015) *Verse: An Introduction to Prosody*, Wiley-Blackwell, Hoboken, NJ, USA.
[20] HOLMES, D. I. (1998) The evolution of stylometry in humanities scholarship. *Literary Linguist. Comput.* **50**, 111–117.
[21] HONNIBAL, M. (2017) SPACY. Available at https://spacy.io.
[22] HUGHES, J. M., FOTI, N. J., KRAKAUER, D. C. & ROCKMORE, D. N. (2012) Quantitative patterns of stylistic influence in the evolution of literature. *Proc. Natl. Acad. Sci. U S A* **109**, 7682–7686.
[23] JACKSON, M. P. (2002) Pause patterns in Shakespeare's verse: canon and chronology. *Literary Linguist. Comput.* **17**, 37–46.
[24] KESSLER, B., NUNBERG, G. & SCHUTZE, H. (1996) Automatic detection of text genre. In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*.
[25] KJELL, B. (1994) Authorship attribution of text samples using neural networks and Bayesian classifiers. In: *Proceedings of the 1994 IEEE International Conference on Systems, Man and Cybernetics*, Vol. 2, pp. 1660–1664.
[26] KULLBACK, S. & LEIBLER, R. A. (1951) On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86.
[27] LAI, S., XU, L., LIU, K. & ZHAO, J. (2015) Recurrent convolutional neural networks for text classification. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI '15)*, pp. 2267–2273.
[28] LAWLER, J. (2006) Punctuation. In: Ken Brown (editor), *Encyclopedia of Language & Linguistics*, 2nd ed., Elsevier, Amsterdam, The Netherlands.
[29] LESNE, A. (2014) Shannon entropy: a rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics. *Math. Struct. Comput. Sci.* **24**, e240311.
[30] LEWIS, T. (1979) Notes on punctuation. In: *The Medusa and the Snail: More Notes of a Biology Watcher*, Viking Press, New York City, NY, USA.
[31] LIN, J. (1991) Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **37**, 145–151.
[32] LYONS, T. (2014) Rough paths, signatures and the modelling of functions on streams. In: *Proceedings of the International Congress of Mathematicians 2014, Korea*. Available at http://www.icm2014.org/download/Proceedings_Volume_IV.pdf.

[33] MENDENHALL, T. C. (1887) The characteristic curves of composition. *Science* **9**, 237–249.
[34] MOSTELLER, F. & WALLACE, D. L. (1964) *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, Reading, MA, USA.
[35] NEAL, T., SUNDARARAJAN, K., FATIMA, A. & WOODARD, D. (2018) Surveying stylometry techniques and applications. *ACM Comput. Surv.* **50**, 86.
[36] NEIDORF, L., KRIEGER, M. S., YAKUBEK, M., CHAUDHURI, P. & DEXTER, J. P. (2019) Large-scale quantitative profiling of the old English verse tradition. *Nat. Hum. Behav.* **3**, 560–567.
[37] NUNBERG, G. (1990) *The Linguistics of Punctuation*, Center for the Study of Language and Information, Stanford, CA, USA.
[38] PARKES, M. B. (editor) (1992) *Pause and Effect: An Introduction to the History of Punctuation in the West*, University of California Press, Berkeley, CA, USA.
[39] PULLUM, G. & HUDDLESTON, R. (2001) *The Cambridge Grammar of the English Language*, Cambridge University Press, The Other Place, UK.
[40] QIAN, C., HE, T. & ZHANG, R. (2017) *Deep Learning Based Authorship Identification*. Class report, Computer Science 224, Stanford University. Available at https://pdfs.semanticscholar.org/ab0e/be094ec0a44fb0013d640b344d8cfd7adc81.pdf?_ga=2.215953495.1190289256.1578845031-6826891.1578845031.
[41] SANTINI, M. (2004) A shallow approach to syntactic feature extraction for genre classification. In: *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*.
[42] SANTINI, M. *State-of-the-Art on Automatic Genre Identification*, Information Technology Research Institute (ITRI) Technical Report Series 04-03, University of Brighton, UK, (2004). Available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.5.7680.
[43] SHANNON, C. E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, 379–423, 623–656.
[44] SHLENS, J. (2014) Notes on Kullback–Leibler divergence and likelihood theory. arXiv:1404.2000.
[45] STAMATATOS, E. (2009) A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Tech.* **60**, 538–556.
[46] STAMOU, C. (2008) Stylochronometry: stylistic development, sequence of composition, and relative dating. *Literary Linguist. Comput.* **23**, 181–199.
[47] TRUSS, L. (2004) *Eats, Shoots and Leaves: The Zero Tolerance Approach to Punctuation*, Profile Books, London, UK.
[48] VIEIRA, D. S., PICOLI, S. & MENDES, R. S. (2018) Robustness of sentence length measures in written texts. *Physica A* **506**, 749–754.
[49] WATSON, C. (2019) *Semicolon: The Past, Present, and Future of a Misunderstood Mark*, Ecco Press, New York, NY, USA.
[50] WHISSELL, C. (1996) Traditional and emotional stylometric analysis of the songs of Beatles Paul McCartney and John Lennon. *Comput. Human.* **30**, 257–265.
[51] YANG, A. C.-C., PENG, C.-K., YIEN, H.-W. AND GOLDBERGER, A. (2003) Information categorization approach to literary authorship disputes. *Physica A* **329**, 473–483.
[52] ZHAO, Y., ZOBEL, J. & VINES, P. (2006) Using relative entropy for authorship attribution. In: *Proceedings of the Third Asia Conference on Information Retrieval Technology (AIRS '06)*, pp. 92–105.

Appendix A Author and genre lists

"`Mr Speaker, I said the honourable Member was a liar it is true and I am sorry for it. The honourable Member may place the punctuation where he pleases.`"

— Attributed to Richard Brinsley Sheridan (1751–1816), responding to a rebuke from the Chair for calling a fellow Member of Parliament a liar.

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

In Table A.1, we list the authors that we use in our study. We order them based on their $\boldsymbol{f}^3$ consistency, where smaller numbers indicate greater consistency. (See equation (3.1) for the definition of 'consistency'.) The author order proceeds down the first column and then down the second column. We structure each row as follows: author name (number of documents by that author in our corpus – testing-set size for our experiment on the full corpus – author $\boldsymbol{f}^3$ consistency in our corpus – author accuracy for the testing set for our experiment with the full set of features). Consistency values that are closer to 0 correspond to authors who are more consistent, and accuracy values that are closer to 1 indicate that we correctly assign a larger fraction of books by that author. (See equation (2.5) for the definition of 'accuracy'.) The designation 'NA' indicates that an author is not in the testing set. We number each row in Table A.1 to facilitate the referencing of specific authors. One number references two distinct authors (with one in each column), and we increment the row number from page to page in a way that accounts for the number of authors in the second column.

In Table A.2, we list the genres that we use in our study. We order them based on their $\boldsymbol{f}^3$ consistency. The genre order proceeds down the first column and then down the second column. We structure each row as follows: genre (number of documents in the genre – testing-set size for our experiment on the full corpus – genre $\boldsymbol{f}^3$ consistency in our corpus – genre accuracy for the testing set for our experiment with the full set of features). Consistency values that are closer to 0 correspond to genres that are more consistent, and accuracy values that are closer to 1 indicate that we correctly assign a larger fraction of books of that genre. We number each row in Table A.2 to facilitate the referencing of specific genres. One number references two distinct genres (with one in each column).

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

Table A.1. *The authors that we use in our study.*

	Author (No. documents - test size - consistency - accuracy)	Author (No. documents - test size - consistency - accuracy)
0	Matthews, Stanley R. (32 - 5 - 0.018 - 1.0)	Werner, E. (18 - 2 - 0.053 - 1.0)
1	Hill, Grace Brooks (11 - 2 - 0.02 - 0.0)	Kyne, Peter B. (Peter Bernard) (10 - 2 - 0.054 - 0.0)
2	Dell, Ethel M. (Ethel May) (16 - 5 - 0.02 - 0.8)	Wood, Henry, Mrs. (24 - 6 - 0.055 - 1.0)
3	Goodwin, Harold L. (Harold Leland) (13 - 2 - 0.021 - 1.0)	King, Charles (27 - 6 - 0.055 - 1.0)
4	Young, Clarence (23 - 7 - 0.023 - 0.857)	Bassett, Sara Ware (16 - 3 - 0.055 - 0.333)
5	Hancock, H. Irving (Harrie Irving) (40 - 9 - 0.024 - 1.0)	Abbott, John S. C. (John Stevens Cabot) (23 - 7 - 0.056 - 1.0)
6	Wirt, Mildred A. (Mildred Augustine) (30 - 5 - 0.024 - 1.0)	Gregory, Jackson (10 - NA - 0.056 - NA)
7	United States. Warren Commission (12 - 1 - 0.025 - 1.0)	Maclaren, Alexander (20 - 7 - 0.056 - 1.0)
8	Merriman, Henry Seton (14 - 3 - 0.026 - 1.0)	De Quincey, Thomas (20 - 3 - 0.056 - 1.0)
9	Brame, Charlotte M. (11 - 1 - 0.026 - 1.0)	Aimard, Gustave (29 - 5 - 0.056 - 0.8)
10	Patchin, Frank Gee (15 - NA - 0.026 - NA)	Mundy, Talbot (13 - 3 - 0.056 - 1.0)
11	Norris, Kathleen Thompson (11 - 3 - 0.027 - 1.0)	Carey, Rosa Nouchette (11 - 2 - 0.056 - 0.5)
12	Hayes, Clair W. (Clair Wallace) (18 - 2 - 0.028 - 1.0)	Barbour, Ralph Henry (32 - 6 - 0.056 - 1.0)
13	Hocking, Joseph (11 - 3 - 0.028 - 0.333)	Goldfrap, John Henry (37 - 7 - 0.056 - 0.857)
14	Locke, William John (21 - 4 - 0.028 - 0.75)	Nicholson, Meredith (13 - 3 - 0.056 - 0.667)
15	Henry, O. (13 - 2 - 0.028 - 1.0)	Tarkington, Booth (19 - 4 - 0.056 - 0.75)
16	Parrish, Randall (15 - 4 - 0.03 - 1.0)	Packard, Frank L. (Frank Lucius) (11 - 2 - 0.057 - 1.0)
17	Bowen, Robert Sidney (15 - 3 - 0.031 - 1.0)	Dowling, Richard (16 - 1 - 0.058 - 0.0)
18	Lynde, Francis (17 - 2 - 0.031 - 1.0)	Ainsworth, William Harrison (20 - 1 - 0.058 - 1.0)
19	Bloundelle-Burton, John (14 - 3 - 0.031 - 0.667)	Everett-Green, Evelyn (19 - 6 - 0.058 - 0.833)
20	Suetonius (14 - 3 - 0.031 - 1.0)	Saint-Simon, Louis de Rouvroy, duc de (15 - 2 - 0.058 - 0.5)
21	Wairy, Louis Constant (12 - 3 - 0.033 - 1.0)	Thorne, Guy (15 - 3 - 0.059 - 0.667)
22	Blanchard, Amy Ella (12 - 3 - 0.033 - 0.0)	Seltzer, Charles Alden (10 - 2 - 0.059 - 0.5)
23	Cholmondeley, Mary (11 - 4 - 0.036 - 0.25)	Meade, L. T. (52 - 12 - 0.059 - 0.667)
24	Buffon, Georges Louis Leclerc, comte de (10 - 1 - 0.036 - 1.0)	Douglas, Amanda M. (19 - 2 - 0.059 - 0.5)
25	Walton, Amy (10 - 2 - 0.036 - 1.0)	Fitzhugh, Percy Keese (22 - 4 - 0.06 - 1.0)
26	Ferber, Edna (10 - 2 - 0.036 - 1.0)	Oppenheim, E. Phillips (Edward Phillips) (58 - 14 - 0.06 - 1.0)
27	Hope, Laura Lee (64 - 11 - 0.037 - 0.818)	Stephens, Ann S. (Ann Sophia) (13 - 5 - 0.06 - 0.6)
28	Chadwick, Lester (16 - 4 - 0.037 - 0.75)	Fyfe, H. B. (Horace Bowne) (16 - 2 - 0.061 - 1.0)
29	Mitford, Bertram (27 - 2 - 0.038 - 1.0)	Wodehouse, P. G. (Pelham Grenville) (37 - 4 - 0.061 - 1.0)
30	Appleton, Victor (31 - 5 - 0.038 - 0.4)	Deland, Margaret Wade Campbell (11 - 3 - 0.061 - 0.667)
31	Penrose, Margaret (22 - 2 - 0.039 - 0.5)	Holt, Emily Sarah (22 - 5 - 0.061 - 0.8)
32	Collingwood, Harry (33 - 6 - 0.039 - 1.0)	Carter, Herbert, active 1909-1917 (12 - NA - 0.061 - NA)
33	Finley, Martha (35 - 8 - 0.04 - 0.625)	Porter, Eleanor H. (Eleanor Hodgman) (13 - 4 - 0.062 - 0.75)
34	Mackintosh, Charles Henry (11 - 2 - 0.04 - 1.0)	Moore, Frank Frankfort (19 - 6 - 0.062 - 1.0)
35	Phillips, David Graham (14 - 2 - 0.04 - 0.5)	Farjeon, B. L. (Benjamin Leopold) (29 - 4 - 0.062 - 1.0)
36	Boldrewood, Rolf (15 - NA - 0.04 - NA)	Snell, Roy J. (Roy Judson) (40 - 10 - 0.062 - 1.0)
37	Harper, Charles G. (Charles George) (16 - NA - 0.04 - NA)	Kock, Paul de (18 - 6 - 0.062 - 0.833)
38	Weyman, Stanley John (28 - 11 - 0.041 - 1.0)	Johnson, Owen (11 - 3 - 0.062 - 0.667)
39	Roy, Lillian Elizabeth (16 - 3 - 0.041 - 1.0)	Walsh, James J. (James Joseph) (12 - 3 - 0.063 - 1.0)
40	Emerson, Alice B. (23 - 2 - 0.042 - 0.0)	Blackwood, Algernon (22 - 6 - 0.063 - 0.667)
41	McCutcheon, George Barr (33 - 6 - 0.043 - 0.667)	Craik, Dinah Maria Mulock (15 - 3 - 0.063 - 0.333)
42	Reeve, Arthur B. (Arthur Benjamin) (14 - NA - 0.044 - NA)	Marlowe, Stephen (16 - 4 - 0.063 - 0.5)
43	Shaler, Robert (18 - 5 - 0.044 - 0.4)	Harben, Will N. (Will Nathaniel) (13 - NA - 0.063 - NA)
44	Bourrienne, Louis Antoine Fauvelet de (16 - 3 - 0.044 - 0.333)	Robertson, Margaret M. (Margaret Murray) (11 - 1 - 0.064 - 1.0)
45	Vaizey, George de Horne, Mrs. (22 - NA - 0.044 - NA)	De Mille, James (17 - 3 - 0.064 - 0.667)
46	Mathews, Joanna H. (Joanna Hooe) (13 - 2 - 0.044 - 1.0)	Rockwood, Roy (16 - 1 - 0.064 - 0.0)
47	Vance, Louis Joseph (12 - NA - 0.045 - NA)	Holmes, Mary Jane (21 - 3 - 0.064 - 1.0)
48	Duncan, Sara Jeannette (10 - 1 - 0.045 - 0.0)	Mühlbach L. (Luise) (20 - 2 - 0.064 - 0.5)
49	Pansy (11 - 1 - 0.045 - 0.0)	Leslie, Madeline (20 - 4 - 0.065 - 1.0)
50	Raine, William MacLeod (22 - 4 - 0.046 - 1.0)	Oliphant, Mrs. (Margaret) (70 - 14 - 0.066 - 0.929)
51	Douglas, Alan, Captain (10 - 4 - 0.046 - 0.75)	Boothby, Guy (16 - 4 - 0.066 - 0.25)
52	MacGrath, Harold (21 - NA - 0.048 - NA)	Green, Anna Katharine (35 - 5 - 0.066 - 0.8)
53	Cannon, Richard (26 - 5 - 0.048 - 1.0)	Williamson, C. N. (Charles Norris) (19 - 5 - 0.066 - 0.4)
54	Warner, Susan (25 - 2 - 0.048 - 1.0)	Hale, Edward Everett (10 - 5 - 0.066 - 0.2)
55	Cody, H. A. (Hiram Alfred) (12 - 3 - 0.048 - 1.0)	Aycock, Roger D. (12 - 4 - 0.066 - 0.75)
56	Brazil, Angela (27 - 4 - 0.048 - 1.0)	Daviess, Maria Thompson (11 - 2 - 0.067 - 1.0)
57	Barr, Robert (20 - 4 - 0.048 - 0.75)	Day, Holman (11 - NA - 0.067 - NA)
58	Rice, Alice Caldwell Hegan (10 - 4 - 0.049 - 0.0)	Chambers, Robert W. (Robert William) (43 - 9 - 0.067 - 0.889)
59	Frey, Hildegard G. (10 - NA - 0.049 - NA)	Munroe, Kirk (15 - 3 - 0.067 - 0.667)
60	Southworth, Emma Dorothy Eliza Nevitte (13 - 2 - 0.049 - 1.0)	Blackmore, R. D. (Richard Doddridge) (23 - 5 - 0.068 - 1.0)
61	Standish, Burt L. (25 - 2 - 0.049 - 1.0)	Mansfield, M. F. (Milburg Francisco) (16 - 4 - 0.068 - 0.75)
62	Tracy, Louis (27 - 5 - 0.049 - 0.6)	Crockett, S. R. (Samuel Rutherford) (19 - 4 - 0.068 - 0.75)
63	Altsheler, Joseph A. (Joseph Alexander) (33 - 8 - 0.049 - 1.0)	Chase, Josephine (32 - 4 - 0.068 - 0.75)
64	Skinner, Charles M. (Charles Montgomery) (10 - 2.0 - 0.05 - 1.0)	Heyse, Paul (10 - 4.0 - 0.068 - 0.25)
65	Hutcheson, John C. (John Conroy) (17 - 1 - 0.05 - 1.0)	Buck, Charles Neville (11 - 1 - 0.068 - 1.0)
66	Braddon, M. E. (Mary Elizabeth) (30 - 5 - 0.05 - 1.0)	Mangasarian, M. M. (Mangasar Mugurditch) (12 - NA - 0.069 - NA)
67	Comstock, Harriet T. (Harriet Theresa) (10 - 4 - 0.051 - 0.5)	Shakespeare (spurious and doubtful works) (10 - 1 - 0.069 - 0.0)
68	Glasgow, Ellen Anderson Gholson (12 - 3 - 0.051 - 0.667)	Riis, Jacob A. (Jacob August) (11 - 2 - 0.069 - 0.0)
69	Beach, Rex (16 - 4 - 0.052 - 0.75)	Miller, Alex. McVeigh, Mrs. (17 - 2 - 0.069 - 1.0)
70	Cullum, Ridgwell (17 - 2 - 0.052 - 1.0)	Westerman, Percy F. (Percy Francis) (34 - 10 - 0.07 - 0.9)
71	Stratemeyer, Edward (75 - 13 - 0.052 - 0.923)	Ewing, Juliana Horatia Gatty (20 - 3 - 0.07 - 0.667)
72	May, Sophie (25 - 2 - 0.052 - 1.0)	Schubin, Ossip (10 - 2 - 0.07 - 0.0)
73	Bower, B. M. (29 - 6 - 0.052 - 1.0)	Lavell, Edith (11 - 1 - 0.071 - 1.0)
74	Fleming, May Agnes (11 - 2 - 0.052 - 0.5)	James, G. P. R. (George Payne Rainsford) (49 - 7 - 0.071 - 1.0)

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

Table A.1. *Continued*

	Author (No. documents - test size - consistency - accuracy)	Author (No. documents - test size - consistency - accuracy)
150	Brandes, Georg (11 - 4 - 0.071 - 0.5)	Haggard, H. Rider (Henry Rider) (51 - 9 - 0.091 - 0.778)
151	Brand, Max (14 - 1 - 0.071 - 1.0)	Jameson, Mrs. (Anna) (10 - NA - 0.091 - NA)
152	Steel, Flora Annie Webster (20 - 6 - 0.071 - 1.0)	Maspero, G. (Gaston) (10 - 3 - 0.091 - 1.0)
153	Smith, E. E. (Edward Elmer) (10 - 3 - 0.072 - 0.667)	Perkins, Lucy Fitch (13 - 4 - 0.091 - 0.5)
154	Garis, Howard Roger (34 - 7 - 0.072 - 0.429)	Schmitz, James H. (10 - NA - 0.091 - NA)
155	Gaboriau, Emile (14 - 3 - 0.072 - 1.0)	Gale, Zona (10 - 3 - 0.091 - 0.667)
156	Smiles, Samuel (14 - 1 - 0.072 - 0.0)	Pater, Walter (13 - 1 - 0.091 - 0.0)
157	Henty, G. A. (George Alfred) (104 - 23 - 0.072 - 0.913)	Holinshed, Raphael (27 - 3 - 0.092 - 1.0)
158	Arthur, T. S. (Timothy Shay) (32 - 10 - 0.074 - 0.6)	Wallace, F. L. (Floyd L.) (13 - 1 - 0.092 - 1.0)
159	Raymond, Evelyn (17 - 3 - 0.074 - 1.0)	Strang, Herbert (32 - 7 - 0.092 - 1.0)
160	Nye, Bill (11 - NA - 0.074 - NA)	Catherwood, Mary Hartwell (20 - 8 - 0.092 - 0.625)
161	James, William (11 - 1 - 0.074 - 1.0)	Norris, Frank (10 - 4 - 0.092 - 0.25)
162	Speed, Nell (16 - 5 - 0.075 - 0.8)	Lincoln, Joseph Crosby (18 - 2 - 0.093 - 1.0)
163	Barr, Amelia E. (26 - 4 - 0.075 - 1.0)	Cawein, Madison Julius (19 - 1 - 0.093 - 1.0)
164	Ashton, John (17 - 4 - 0.075 - 0.75)	Alcott, Louisa May (37 - 5 - 0.094 - 0.8)
165	Mill, John Stuart (14 - 2 - 0.075 - 1.0)	Walpole, Hugh (12 - 2 - 0.094 - 0.5)
166	Ellis, Havelock (12 - 2 - 0.076 - 1.0)	Sharp, Dallas Lore (10 - 3 - 0.094 - 1.0)
167	Stephens, Robert Neilson (10 - 4 - 0.076 - 0.25)	Pepys, Samuel (76 - 18 - 0.095 - 1.0)
168	Harmon, Jim (13 - 4 - 0.077 - 0.5)	Harland, Henry (12 - 5 - 0.095 - 0.8)
169	Oxley, J. Macdonald (James Macdonald) (10 - 1 - 0.077 - 1.0)	Fox, John (13 - 3 - 0.095 - 0.667)
170	Marryat, Frederick (36 - 6 - 0.077 - 1.0)	Black, William (20 - 5 - 0.096 - 0.8)
171	Hendryx, James B. (James Beardsley) (10 - 1 - 0.077 - 1.0)	Thoreau, Henry David (11 - 2 - 0.096 - 0.0)
172	Jefferson, Thomas (17 - 6 - 0.077 - 1.0)	Smith, George O. (George Oliver) (10 - 1 - 0.096 - 1.0)
173	Cory, David (15 - 3 - 0.077 - 0.667)	Lord, John (18 - 5 - 0.097 - 0.6)
174	Casanova, Giacomo (32 - 5 - 0.078 - 0.8)	Burroughs, John (23 - 5 - 0.097 - 1.0)
175	James, George Wharton (11 - 2 - 0.078 - 0.0)	Turgenev, Ivan Sergeevich (22 - 5 - 0.097 - 0.6)
176	Bindloss, Harold (43 - 11 - 0.078 - 1.0)	Ballantyne, R. M. (Robert Michael) (91 - 21 - 0.098 - 0.952)
177	Sheckley, Robert (18 - 6 - 0.078 - 0.667)	Spencer, Herbert (10 - 1 - 0.098 - 1.0)
178	Hope, Anthony (33 - 5 - 0.078 - 0.6)	Lowndes, Marie Belloc (15 - NA - 0.098 - NA)
179	Bailey, Arthur Scott (40 - 10 - 0.079 - 1.0)	Buchanan, Robert Williams (10 - 1 - 0.098 - 1.0)
180	Loti, Pierre (11 - 3 - 0.079 - 0.667)	Maupassant, Guy de (33 - 8 - 0.098 - 0.75)
181	Vandercook, Margaret (24 - 4 - 0.079 - 1.0)	Stacpoole, H. De Vere (Henry De Vere) (20 - 6 - 0.099 - 0.167)
182	Senarens, Luis (15 - 2 - 0.08 - 1.0)	Crane, Stephen (13 - 3 - 0.099 - 0.333)
183	Sedgwick, Anne Douglas (14 - 2 - 0.08 - 0.5)	Murfree, Mary Noailles (26 - 2 - 0.099 - 0.5)
184	Whyte-Melville, G. J. (George John) (10 - 5 - 0.08 - 0.0)	Jókai, Mór (28 - 9 - 0.099 - 0.444)
185	Williamson, A. M. (Alice Muriel) (15 - 3 - 0.08 - 0.333)	Ouida (22 - 2 - 0.1 - 1.0)
186	Burgess, Thornton W. (Thornton Waldo) (37 - 8 - 0.081 - 1.0)	Moody, Dwight Lyman (14 - 2 - 0.1 - 1.0)
187	Dick, Philip K. (12 - 1 - 0.081 - 0.0)	Auerbach, Berthold (10 - 2 - 0.101 - 0.5)
188	Harris, Frank (10 - NA - 0.081 - NA)	Garland, Hamlin (23 - 2 - 0.101 - 1.0)
189	Von Arnim, Elizabeth (12 - 2 - 0.082 - 1.0)	Smith, Evelyn E. (15 - 3 - 0.101 - 1.0)
190	Grey, Zane (26 - 4 - 0.082 - 1.0)	Hay, Ian (13 - 1 - 0.102 - 1.0)
191	Fenn, George Manville (128 - 28 - 0.082 - 0.964)	Garrett, Randall (43 - 10 - 0.102 - 0.6)
192	King, Basil (10 - 4 - 0.082 - 1.0)	Stoddard, William Osborn (12 - 2 - 0.102 - 1.0)
193	Castlemon, Harry (38 - 8 - 0.083 - 0.875)	Marsh, Richard (19 - 5 - 0.103 - 0.4)
194	Ward, Humphry, Mrs. (33 - 8 - 0.083 - 0.625)	Ford, Sewell (12 - 3 - 0.103 - 1.0)
195	Burke, Edmund (15 - 3 - 0.084 - 1.0)	Gissing, George (24 - 9 - 0.103 - 0.667)
196	Brereton, F. S. (Frederick Sadleir) (18 - 5 - 0.084 - 0.6)	Leblanc, Maurice (16 - 6 - 0.103 - 1.0)
197	Connor, Ralph (14 - 4 - 0.084 - 1.0)	Bensusan, S. L. (Samuel Levy) (11 - 3 - 0.104 - 0.333)
198	United States. Central Intelligence Agency (21 - 4 - 0.084 - 1.0)	Motley, John Lothrop (89 - 17 - 0.104 - 0.882)
199	Onions, Oliver (11 - 3 - 0.084 - 0.667)	Alger, Horatio, Jr. (95 - 21 - 0.104 - 0.857)
200	Hill, Grace Livingston (15 - 4 - 0.085 - 0.5)	Smith, Francis Hopkinson (26 - 3 - 0.104 - 0.667)
201	Rohmer, Sax (17 - 1 - 0.085 - 0.0)	Mitton, G. E. (Geraldine Edith) (12 - 2 - 0.105 - 0.0)
202	Habberton, John (11 - 2 - 0.085 - 0.5)	Le Queux, William (66 - 8 - 0.106 - 0.875)
203	Russell, William Clark (18 - 10 - 0.085 - 0.4)	Norton, Andre (14 - 5 - 0.106 - 0.6)
204	Richmond, Grace S. (Grace Smith) (15 - 4 - 0.086 - 0.5)	Symonds, John Addington (15 - 2 - 0.106 - 0.5)
205	Bacon, Josephine Daskam (13 - 1 - 0.086 - 1.0)	Santayana, George (10 - 2 - 0.106 - 0.0)
206	Hume, Fergus (63 - 17 - 0.086 - 0.941)	Parkman, Francis (15 - 2 - 0.107 - 0.5)
207	Kingston, William Henry Giles (131 - 32 - 0.086 - 0.938)	Cable, George Washington (14 - 1 - 0.107 - 1.0)
208	Sue, Eugène (44 - 11 - 0.086 - 0.818)	Irving, Washington (20 - 3 - 0.107 - 0.667)
209	Hornung, E. W. (Ernest William) (26 - 2 - 0.086 - 1.0)	MacGregor, Mary Esther Miller (10 - 3 - 0.108 - 0.0)
210	Orczy, Emmuska Orczy, Baroness (18 - 3 - 0.087 - 1.0)	Kjelgaard, Jim (11 - 4 - 0.109 - 0.75)
211	Hulbert, Archer Butler (17 - 1 - 0.087 - 1.0)	Crawford, F. Marion (Francis Marion) (47 - 11 - 0.109 - 0.818)
212	Machen, Arthur (10 - 3 - 0.087 - 0.333)	Romanes, George John (11 - 3 - 0.109 - 1.0)
213	Chapman, Allen (25 - 2 - 0.087 - 0.5)	Farnol, Jeffery (14 - 2 - 0.11 - 1.0)
214	Vasari, Giorgio (11 - 1 - 0.088 - 1.0)	Webster, Frank V. (19 - 3 - 0.111 - 0.333)
215	Mulford, Clarence Edward (10 - 3 - 0.088 - 1.0)	Richards, Laura Elizabeth Howe (42 - 6 - 0.111 - 0.833)
216	Wood, William Charles Henry (12 - NA - 0.088 - NA)	Fiske, John (18 - 4 - 0.113 - 0.5)
217	Mitford, Mary Russell (13 - 1 - 0.089 - 1.0)	Spyri, Johanna (15 - 4 - 0.113 - 1.0)
218	Saunders, Marshall (13 - 2 - 0.089 - 0.0)	Adams, Samuel Hopkins (13 - 2 - 0.114 - 0.5)
219	Frederic, Harold (14 - 2 - 0.089 - 0.5)	Mahan, A. T. (Alfred Thayer) (15 - 2 - 0.114 - 0.5)
220	Molesworth, Mrs. (55 - 8 - 0.089 - 1.0)	Ingersoll, Robert Green (30 - 6 - 0.115 - 1.0)
221	Del Rey, Lester (12 - NA - 0.089 - NA)	Beerbohm, Max, Sir (10 - 3 - 0.115 - 0.333)
222	Macaulay, Thomas Babington Macaulay, Baron (19 - 1 - 0.09 - 0.0)	Hume, David (13 - 2 - 0.115 - 1.0)
223	Reynolds, Mack (24 - 5 - 0.09 - 1.0)	Couperus, Louis (13 - 1 - 0.115 - 1.0)
224	Fletcher, J. S. (Joseph Smith) (17 - 2 - 0.09 - 1.0)	Pemberton, Max (11 - 3 - 0.115 - 0.333)

Table A.1. *Continued*

	Author (No. documents - test size - consistency - accuracy)	Author (No. documents - test size - consistency - accuracy)
300	Rathborne, St. George (14 - 2 - 0.116 - 0.5)	Thackeray, William Makepeace (35 - 7 - 0.148 - 0.571)
301	Cervantes Saavedra, Miguel de (47 - 13 - 0.116 - 0.462)	Trollope, Anthony (78 - 24 - 0.148 - 0.75)
302	Samachson, Joseph (12 - 1 - 0.116 - 0.0)	Howard, Robert E. (Robert Ervin) (12 - 4 - 0.149 - 1.0)
303	Curwood, James Oliver (27 - NA - 0.116 - NA)	Collins, Wilkie (35 - 5 - 0.149 - 0.8)
304	Saintsbury, George (12 - 3 - 0.116 - 0.667)	Pyle, Howard (16 - 1 - 0.149 - 0.0)
305	Johnston, Annie F. (Annie Fellows) (37 - 7 - 0.117 - 0.571)	Hichens, Robert (27 - 5 - 0.15 - 1.0)
306	Gaskell, Elizabeth Cleghorn (23 - 1 - 0.117 - 1.0)	Froude, James Anthony (12 - 4 - 0.15 - 0.75)
307	Nourse, Alan Edward (23 - 3 - 0.117 - 0.667)	Benson, E. F. (Edward Frederic) (28 - 8 - 0.15 - 0.875)
308	Stables, Gordon (26 - 6 - 0.118 - 0.5)	Rinehart, Mary Roberts (29 - 6 - 0.152 - 0.333)
309	Laumer, Keith (12 - 3 - 0.118 - 0.667)	Hurll, Estelle M. (Estelle May) (13 - 4 - 0.152 - 1.0)
310	Hoare, Edward (32 - 8 - 0.118 - 0.5)	Blasco Ibáñez, Vicente (14 - 2 - 0.152 - 1.0)
311	Roe, Edward Payson (19 - 5 - 0.119 - 1.0)	Coolidge, Susan (14 - 4 - 0.152 - 0.75)
312	Bellamy, Edward (20 - 4 - 0.119 - 0.25)	Corelli, Marie (14 - 3 - 0.153 - 0.333)
313	Merwin, Samuel (13 - 3 - 0.119 - 0.333)	Swift, Jonathan (16 - 1 - 0.154 - 0.0)
314	Grant, James, archaeologist (12 - 2 - 0.12 - 1.0)	Dostoyevsky, Fyodor (11 - 2 - 0.155 - 1.0)
315	Wiggin, Kate Douglas Smith (33 - 6 - 0.12 - 0.667)	Tapper, Thomas (13 - 3 - 0.158 - 1.0)
316	Russell, Bertrand (11 - 3 - 0.122 - 0.667)	Burney, Fanny (14 - 2 - 0.158 - 1.0)
317	Ritchie, J. Ewing (James Ewing) (20 - 2 - 0.122 - 0.0)	Willis, Nathaniel Parker (10 - 1 - 0.159 - 1.0)
318	Harrison, Harry (10 - 3 - 0.122 - 0.667)	Optic, Oliver (59 - 16 - 0.16 - 0.812)
319	Müller, F. Max (Friedrich Max) (10 - 2 - 0.122 - 0.5)	Hawthorne, Julian (12 - 3 - 0.16 - 0.0)
320	Dewey, John (15 - 1 - 0.123 - 1.0)	Lever, Charles James (53 - 12 - 0.16 - 0.75)
321	Parker, Gilbert (106 - 18 - 0.124 - 0.778)	Murray, David Christie (14 - 2 - 0.161 - 1.0)
322	Sabatini, Rafael (18 - 4 - 0.124 - 1.0)	Benson, Arthur Christopher (16 - 3 - 0.163 - 0.667)
323	Church, Alfred John (12 - 3 - 0.125 - 0.0)	Wade, Mary Hazelton Blanchard (21 - 4 - 0.164 - 0.75)
324	Marks, Winston K. (12 - 3 - 0.125 - 0.333)	Le Gallienne, Richard (17 - 4 - 0.164 - 0.5)
325	Huneker, James (11 - 1 - 0.125 - 0.0)	Benson, Robert Hugh (11 - 3 - 0.165 - 0.667)
326	Morris, Charles (18 - 4 - 0.127 - 1.0)	Whittier, John Greenleaf (37 - 5 - 0.165 - 1.0)
327	Follen, Eliza Lee Cabot (10 - 4 - 0.127 - 0.5)	Chesterfield, Philip Dormer Stanhope, Earl of (12 - 2 - 0.166 - 1.0)
328	Foote, G. W. (George William) (10 - 1 - 0.127 - 1.0)	Sienkiewicz, Henryk (18 - 3 - 0.167 - 0.0)
329	Harte, Bret (57 - 12 - 0.128 - 0.75)	Gautier, Théophile (11 - 1 - 0.167 - 0.0)
330	Doctorow, Cory (13 - 4 - 0.128 - 1.0)	Wright, Harold Bell (10 - 1 - 0.167 - 0.0)
331	Erckmann-Chatrian (10 - 5 - 0.129 - 0.4)	Scott, Walter (56 - 9 - 0.168 - 0.778)
332	Reid, Mayne (50 - 11 - 0.129 - 0.727)	Frazer, James George (17 - 6 - 0.168 - 1.0)
333	Reed, Talbot Baines (16 - 4 - 0.13 - 0.75)	Doyle, Arthur Conan (61 - 13 - 0.168 - 1.0)
334	Edgeworth, Maria (18 - 6 - 0.13 - 0.5)	Jacobs, W. W. (William Wymark) (105 - 30 - 0.168 - 0.967)
335	Butler, Samuel (18 - 1 - 0.13 - 0.0)	Woolson, Constance Fenimore (14 - 3 - 0.168 - 0.667)
336	Stephen, Leslie (11 - 2 - 0.13 - 1.0)	Roosevelt, Theodore (17 - 1 - 0.169 - 0.0)
337	Piper, H. Beam (33 - 6 - 0.131 - 1.0)	White, Stewart Edward (23 - 8 - 0.17 - 0.875)
338	Hardy, Thomas (26 - 4 - 0.131 - 0.75)	Moodie, Susanna (14 - 2 - 0.17 - 0.5)
339	Dante Alighieri (32 - 5 - 0.132 - 0.6)	Stowe, Harriet Beecher (31 - 4 - 0.171 - 0.25)
340	Cooper, James Fenimore (38 - 4 - 0.132 - 0.5)	Le Fanu, Joseph Sheridan (31 - 9 - 0.172 - 0.667)
341	Baum, L. Frank (Lyman Frank) (54 - 8 - 0.132 - 0.875)	Lee, Vernon (15 - 4 - 0.172 - 0.5)
342	Ruskin, John (47 - 13 - 0.133 - 0.538)	Moore, George Augustus (16 - 2 - 0.173 - 0.5)
343	Roberts, B. H. (Brigham Henry) (14 - 2 - 0.133 - 0.5)	Mason, A. E. W. (Alfred Edward Woodley) (20 - 5 - 0.173 - 0.6)
344	Huxley, Thomas Henry (48 - 13 - 0.133 - 0.692)	Ellis, Edward Sylvester (52 - 10 - 0.173 - 0.9)
345	Birmingham, George A. (15 - 3 - 0.133 - 0.667)	Conwell, Russell H. (11 - 4 - 0.173 - 0.25)
346	Holley, Marietta (16 - 4 - 0.133 - 0.75)	Hough, Emerson (25 - 4 - 0.173 - 0.25)
347	Sinclair, May (21 - 6 - 0.133 - 1.0)	Glyn, Elinor (17 - 5 - 0.175 - 0.6)
348	Lamb, Charles (10 - 3 - 0.133 - 0.667)	Pinero, Arthur Wing (13 - 3 - 0.177 - 0.667)
349	Schopenhauer, Arthur (12 - 1 - 0.134 - 1.0)	Reed, Helen Leah (10 - NA - 0.178 - NA)
350	Atherton, Gertrude Franklin Horn (25 - 5 - 0.135 - 0.8)	Pyle, Katharine (11 - 3 - 0.179 - 1.0)
351	Peattie, Elia Wilkinson (10 - 2 - 0.135 - 0.0)	Caine, Hall, Sir (17 - 2 - 0.18 - 0.5)
352	Atkinson, William Walker (19 - 2 - 0.135 - 1.0)	Wallace, Alfred Russel (13 - 2 - 0.18 - 0.5)
353	Fanny, Aunt (13 - 5 - 0.135 - 0.6)	Eliot, George (15 - 4 - 0.182 - 0.75)
354	Duncan, Norman (10 - 2 - 0.136 - 0.5)	Martineau, Harriet (16 - 1 - 0.182 - 0.0)
355	Ebers, Georg (144 - 28 - 0.136 - 1.0)	MacDonald, George (60 - 12 - 0.182 - 0.75)
356	Morley, John (30 - 4 - 0.137 - 1.0)	Kingsley, Charles (45 - 4 - 0.183 - 1.0)
357	Andersen, H. C. (Hans Christian) (14 - 3 - 0.139 - 0.667)	Cobb, Irvin S. (Irvin Shrewsbury) (24 - 8 - 0.183 - 1.0)
358	Spence, Lewis (10 - 1 - 0.139 - 1.0)	Meynell, Alice (11 - 3 - 0.184 - 0.333)
359	Richardson, Samuel (14 - 4 - 0.141 - 0.75)	Bates, Arlo (14 - 5 - 0.184 - 0.6)
360	Buchan, John (11 - 3 - 0.141 - 0.0)	Montaigne, Michel de (21 - 3 - 0.184 - 1.0)
361	Ralphson, G. Harvey (George Harvey) (14 - 5 - 0.142 - 0.8)	Becke, Louis (39 - 8 - 0.185 - 0.75)
362	Melville, Herman (16 - 4 - 0.142 - 0.5)	Hudson, W. H. (William Henry) (16 - 3 - 0.185 - 0.667)
363	Quiller-Couch, Arthur (40 - 7 - 0.143 - 0.571)	Wordsworth, William (14 - NA - 0.185 - NA)
364	Euripides (10 - 2 - 0.144 - 1.0)	Stockton, Frank Richard (33 - 9 - 0.185 - 0.778)
365	Griffiths, Arthur (18 - 2 - 0.144 - 0.5)	Zangwill, Israel (15 - 4 - 0.186 - 0.5)
366	Carlyle, Thomas (35 - 10 - 0.144 - 0.9)	Jefferies, Richard (20 - 5 - 0.187 - 0.6)
367	Singmaster, Elsie (11 - 2 - 0.145 - 1.0)	Hergesheimer, Joseph (13 - 4 - 0.187 - 1.0)
368	Lytton, Edward Bulwer Lytton, Baron (194 - 43 - 0.145 - 0.93)	MacKenzie, Compton (12 - 2 - 0.188 - 0.0)
369	Abbott, Eleanor Hallowell (10 - 4 - 0.145 - 0.75)	Beers, Henry A. (Henry Augustin) (10 - 2 - 0.188 - 0.0)
370	Grinnell, George Bird (13 - 2 - 0.146 - 1.0)	Hewlett, Maurice (15 - 3 - 0.189 - 0.0)
371	De la Mare, Walter (10 - 3 - 0.146 - 0.333)	Churchill, Winston (62 - 11 - 0.189 - 0.818)
372	Allen, James Lane (13 - 6 - 0.146 - 0.333)	Sharkey, Jack (10 - 1 - 0.189 - 1.0)
373	Bacheller, Irving (18 - 5 - 0.147 - 0.6)	Leiber, Fritz (22 - 6 - 0.191 - 0.5)
374	Emerson, Ralph Waldo (12 - 3 - 0.147 - 0.667)	Flaubert, Gustave (14 - 5 - 0.191 - 0.6)

Table A.1. *Continued*

	Author (No. documents - test size - consistency - accuracy)	Author (No. documents - test size - consistency - accuracy)
450	Hegel, Georg Wilhelm Friedrich (10 - 2 - 0.192 - 1.0)	Molière (20 - 4 - 0.244 - 0.75)
451	Daudet, Alphonse (17 - 3 - 0.193 - 0.333)	Fletcher, John (15 - 4 - 0.244 - 0.0)
452	Brinton, Daniel G. (Daniel Garrison) (18 - 3 - 0.193 - 0.667)	Lebert, Marie (15 - 4 - 0.247 - 0.75)
453	France, Anatole (31 - 3 - 0.194 - 0.667)	Schoolcraft, Henry Rowe (13 - 3 - 0.247 - 0.333)
454	Hakluyt, Richard (15 - 3 - 0.195 - 1.0)	Saltus, Edgar (13 - 4 - 0.249 - 0.25)
455	Duellman, William Edward (12 - 2 - 0.195 - 0.5)	Ballou, Maturin Murray (19 - 5 - 0.249 - 0.4)
456	Janifer, Laurence M. (12 - 2 - 0.196 - 1.0)	Page, Thomas Nelson (24 - 6 - 0.25 - 0.5)
457	Lincoln, Abraham (19 - 5 - 0.196 - 0.2)	Hall, E. Raymond (Eugene Raymond) (15 - 3 - 0.252 - 0.667)
458	Franklin, Benjamin (10 - 3 - 0.198 - 0.333)	Meredith, George (94 - 27 - 0.252 - 0.889)
459	Leacock, Stephen (14 - 2 - 0.199 - 0.0)	Moore, Thomas (12 - 2 - 0.255 - 0.0)
460	Guiney, Louise Imogen (13 - NA - 0.199 - NA)	Janvier, Thomas A. (Thomas Allibone) (13 - 3 - 0.257 - 0.333)
461	Jonson, Ben (12 - 1 - 0.199 - 0.0)	Potter, Beatrix (21 - 5 - 0.257 - 1.0)
462	Zola, Émile (37 - 11 - 0.199 - 0.818)	Wallace, Edgar (16 - 5 - 0.257 - 0.6)
463	Warner, Charles Dudley (41 - 10 - 0.2 - 0.3)	Boswell, James (12 - 3 - 0.258 - 0.667)
464	Cabell, James Branch (13 - 3 - 0.2 - 0.667)	Harris, Joel Chandler (14 - 1 - 0.258 - 0.0)
465	Burton, Richard Francis, Sir (20 - 6 - 0.2 - 0.833)	Young, Filson (11 - 3 - 0.26 - 0.333)
466	Dawson, Coningsby (15 - 2 - 0.201 - 1.0)	Grote, George (13 - 3 - 0.26 - 1.0)
467	Seton, Ernest Thompson (15 - 2 - 0.201 - 0.5)	Allen, Grant (29 - 4 - 0.263 - 0.25)
468	Reade, Charles (15 - 2 - 0.201 - 1.0)	Bone, Jesse F. (Jesse Franklin) (12 - 1 - 0.264 - 1.0)
469	Beaumont, Francis (10 - 2 - 0.202 - 0.5)	Harland, Marion (13 - 3 - 0.266 - 0.667)
470	Bierce, Ambrose (17 - 4 - 0.202 - 0.25)	Phillpotts, Eden (19 - 3 - 0.268 - 0.333)
471	Aldrich, Thomas Bailey (19 - 5 - 0.203 - 0.0)	James, Henry (75 - 10 - 0.269 - 1.0)
472	Rousseau, Jean-Jacques (18 - 5 - 0.203 - 0.4)	Wallace, Dillon (11 - 3 - 0.272 - 0.333)
473	Wharton, Edith (33 - 10 - 0.204 - 0.8)	Borrow, George (39 - 2 - 0.273 - 1.0)
474	Yonge, Charlotte M. (Charlotte Mary) (59 - 8 - 0.204 - 1.0)	Byron, George Gordon Byron, Baron (12 - 3 - 0.273 - 0.667)
475	Bunyan, John (14 - 2 - 0.205 - 0.0)	Mitchell, S. Weir (Silas Weir) (12 - 2 - 0.274 - 0.0)
476	Browning, Robert (10 - 1 - 0.205 - 1.0)	Mencken, H. L. (Henry Louis) (10 - 2 - 0.275 - 0.5)
477	Dryden, John (20 - 5 - 0.206 - 0.8)	Plato (27 - 3 - 0.275 - 1.0)
478	Hubbard, Elbert (20 - 3 - 0.206 - 1.0)	Weymouth, Richard Francis (25 - 4 - 0.275 - 1.0)
479	Hearn, Lafcadio (22 - 7 - 0.207 - 0.429)	Lewis, Alfred Henry (15 - 4 - 0.278 - 0.75)
480	Paine, Albert Bigelow (29 - 6 - 0.208 - 0.667)	Disraeli, Benjamin, Earl of Beaconsfield (17 - 1 - 0.279 - 0.0)
481	Roberts, Charles G. D., Sir (26 - 4 - 0.208 - 1.0)	Eggleston, Edward (12 - 1 - 0.28 - 0.0)
482	Baring-Gould, S. (Sabine) (57 - 9 - 0.208 - 0.667)	Baldwin, James (11 - 1 - 0.283 - 0.0)
483	Freeman, Mary Eleanor Wilkins (23 - 5 - 0.21 - 0.4)	Besant, Walter (19 - 3 - 0.283 - 0.333)
484	Twain, Mark (142 - 32 - 0.21 - 0.812)	Walpole, Horace (12 - 4 - 0.287 - 0.5)
485	Davis, Richard Harding (49 - 9 - 0.21 - 0.667)	Laut, Agnes C. (12 - 1 - 0.288 - 0.0)
486	Verne, Jules (46 - 13 - 0.212 - 0.923)	Stevenson, Robert Louis (70 - 14 - 0.29 - 0.857)
487	Leland, Charles Godfrey (10 - 2 - 0.212 - 0.0)	Carleton, William (21 - 4 - 0.29 - 1.0)
488	Dixon, Thomas (13 - 4 - 0.213 - 1.0)	Wells, H. G. (Herbert George) (51 - 12 - 0.293 - 0.75)
489	Besant, Annie (17 - 1 - 0.214 - 1.0)	Conrad, Joseph (31 - 3 - 0.295 - 1.0)
490	Hawthorne, Nathaniel (92 - 22 - 0.215 - 0.864)	Van Dyke, Henry (29 - 7 - 0.296 - 0.571)
491	Bangs, John Kendrick (37 - 8 - 0.216 - 0.875)	Herford, Oliver (13 - 3 - 0.296 - 0.0)
492	Coleridge, Samuel Taylor (18 - 2 - 0.217 - 1.0)	Herrick, Robert (11 - 3 - 0.297 - 0.0)
493	Voltaire (19 - 6 - 0.218 - 0.833)	Morris, William (28 - 12 - 0.298 - 0.25)
494	Maclaren, Ian (13 - 5 - 0.218 - 0.8)	Adams, Andy (10 - 2 - 0.301 - 1.0)
495	Dickens, Charles (79 - 15 - 0.218 - 0.667)	Marlowe, Christopher (10 - 3 - 0.305 - 0.667)
496	Wells, Carolyn (58 - 16 - 0.219 - 0.688)	Chesterton, G. K. (Gilbert Keith) (37 - 9 - 0.305 - 0.667)
497	Eggleston, George Cary (17 - 3 - 0.221 - 0.0)	Chekhov, Anton Pavlovich (23 - 5 - 0.306 - 0.8)
498	Hughes, Rupert (12 - 1 - 0.221 - 1.0)	London, Jack (50 - 11 - 0.306 - 0.818)
499	Nesbit, E. (Edith) (30 - 6 - 0.224 - 0.5)	Wilson, Harry Leon (13 - 1 - 0.306 - 0.0)
500	Lucas, E. V. (Edward Verrall) (11 - 4 - 0.224 - 0.25)	Wilcox, Ella Wheeler (23 - 8 - 0.306 - 0.5)
501	Hugo, Victor (15 - 6 - 0.224 - 0.833)	Shakespeare, William (105 - 19 - 0.307 - 0.842)
502	Field, Eugene (14 - 2 - 0.227 - 1.0)	Fielding, Henry (14 - 5 - 0.308 - 0.2)
503	Defoe, Daniel (44 - 11 - 0.23 - 0.545)	Phelps, Elizabeth Stuart (14 - 4 - 0.31 - 0.0)
504	Belloc, Hilaire (27 - 5 - 0.231 - 0.8)	‘Abdu’l-Bahá (15 - NA - 0.314 - NA)
505	Darwin, Charles (30 - 2 - 0.235 - 0.5)	Graham, Harry (10 - 2 - 0.319 - 0.5)
506	Drake, Samuel Adams (10 - 1 - 0.235 - 0.0)	Tagore, Rabindranath (19 - 4 - 0.322 - 0.0)
507	Cicero, Marcus Tullius (14 - 3 - 0.235 - 0.0)	Webster, Jean (10 - 1 - 0.325 - 1.0)
508	Newman, John Henry (14 - 1 - 0.236 - 1.0)	Masefield, John (17 - 2 - 0.327 - 1.0)
509	Balzac, Honoré de (119 - 18 - 0.236 - 0.833)	Longfellow, Henry Wadsworth (14 - 4 - 0.327 - 0.25)
510	Butler, Ellis Parker (22 - 3 - 0.236 - 0.333)	Otis, James (45 - 9 - 0.33 - 0.889)
511	Johnston, Mary (18 - 2 - 0.236 - 0.5)	Burroughs, Edgar Rice (19 - 1 - 0.335 - 1.0)
512	Leinster, Murray (37 - 9 - 0.236 - 0.778)	Haeckel, Ernst (13 - 6 - 0.337 - 0.333)
513	O’Donnell, Elliott (10 - 2 - 0.237 - 0.0)	Johnson, Samuel (23 - 6 - 0.337 - 0.5)
514	Wister, Owen (13 - 4 - 0.237 - 0.0)	Jewett, Sarah Orne (12 - 2 - 0.339 - 0.5)
515	McElroy, John (15 - NA - 0.238 - NA)	Luther, Martin (18 - 4 - 0.34 - 0.0)
516	United States. Work Projects Administration (34 - 6 - 0.239 - 1.0)	Homer (12 - 5 - 0.341 - 0.0)
517	La Fontaine, Jean de (31 - 6 - 0.242 - 0.5)	Warner, Anne (10 - 2 - 0.35 - 0.0)
518	Lang, Andrew (72 - 17 - 0.242 - 0.882)	Bennett, Arnold (44 - 16 - 0.35 - 0.875)
519	Brady, Cyrus Townsend (13 - 4 - 0.242 - 0.0)	Home, Gordon (15 - 5 - 0.351 - 0.4)
520	Burnett, Frances Hodgson (41 - 6 - 0.242 - 0.667)	Nietzsche, Friedrich Wilhelm (17 - 1 - 0.351 - 1.0)
521	Dumas, Alexandre (58 - 10 - 0.243 - 0.8)	Abbott, Jacob (51 - 11 - 0.359 - 0.727)
522	Gibbon, Edward (11 - 1 - 0.243 - 1.0)	Gibbs, George (15 - 3 - 0.364 - 1.0)
523	Duchess (16 - 1 - 0.244 - 1.0)	Baker, George M. (George Melville) (19 - 4 - 0.365 - 1.0)
524	Eddy, Mary Baker (10 - 2 - 0.244 - 0.5)	Rolland, Romain (12 - 3 - 0.365 - 0.333)

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press

Table A.1. *Continued*

	Author (No. documents - test size - consistency - accuracy)
600	Jackson, Helen Hunt (13 - 2 - 0.369 - 0.0)
601	Crane, Walter (17 - 4 - 0.37 - 0.5)
602	Goethe, Johann Wolfgang von (15 - 2 - 0.371 - 0.5)
603	Bjørnson, Bjørnstjerne (16 - 3 - 0.372 - 1.0)
604	Carroll, Lewis (19 - 4 - 0.373 - 1.0)
605	Kipling, Rudyard (44 - 10 - 0.373 - 0.7)
606	Riley, James Whitcomb (17 - 2 - 0.377 - 0.5)
607	Jerome, Jerome K. (Jerome Klapka) (32 - 6 - 0.379 - 0.333)
608	Stevenson, Burton Egbert (17 - 4 - 0.382 - 0.25)
609	Webster, Noah (11 - 1 - 0.388 - 1.0)
610	Gorky, Maksim (10 - 1 - 0.393 - 0.0)
611	Peck, George W. (George Wilbur) (10 - 2 - 0.395 - 1.0)
612	Howells, William Dean (94 - 23 - 0.4 - 0.783)
613	Stringer, Arthur (10 - NA - 0.402 - NA)
614	Andreyev, Leonid (11 - NA - 0.403 - NA)
615	Xenophon (16 - 3 - 0.405 - 0.667)
616	Swinburne, Algernon Charles (25 - 2 - 0.406 - 0.5)
617	Yeats, W. B. (William Butler) (35 - 5 - 0.414 - 0.4)
618	Ibsen, Henrik (18 - 4 - 0.415 - 0.25)
619	Montgomery, L. M. (Lucy Maud) (12 - 2 - 0.416 - 1.0)
620	Library of Congress. Copyright Office (66 - 13 - 0.425 - 0.923)
621	Wilson, Ann (12 - 2 - 0.426 - 1.0)
622	Morley, Christopher (12 - 2 - 0.428 - 1.0)
623	Galsworthy, John (47 - 9 - 0.437 - 1.0)
624	Tennyson, Alfred Tennyson, Baron (12 - 2 - 0.44 - 0.5)
625	Shoghi, Effendi (17 - 5 - 0.445 - 1.0)
626	Reed, Myrtle (13 - 3 - 0.451 - 0.333)
627	Holmes, Oliver Wendell (33 - 10 - 0.462 - 0.6)
628	Lawrence, D. H. (David Herbert) (20 - 6 - 0.473 - 0.667)
629	Shaw, Bernard (42 - 8 - 0.481 - 0.75)
630	Anstey, F. (18 - 2 - 0.493 - 1.0)
631	Strindberg, August (22 - 4 - 0.505 - 0.75)
632	Bahá'u'lláh (11 - 5 - 0.508 - 0.2)
633	Burgess, Gelett (11 - 2 - 0.515 - 0.0)
634	Tolstoy, Leo, graf (38 - 8 - 0.521 - 0.75)
635	Bridges, Robert (11 - 2 - 0.523 - 0.5)
636	Spinoza, Benedictus de (12 - 3 - 0.525 - 0.333)
637	Wilde, Oscar (25 - 2 - 0.536 - 0.5)
638	Poe, Edgar Allan (16 - 2 - 0.541 - 0.5)
639	Rice, Cale Young (11 - 3 - 0.544 - 0.333)
640	Barrie, J. M. (James Matthew) (25 - 1 - 0.55 - 1.0)
641	Dunsany, Lord (16 - 5 - 0.575 - 0.2)
642	Maeterlinck, Maurice (18 - 2 - 0.61 - 1.0)
643	Sinclair, Upton (24 - 10 - 0.653 - 0.5)
644	Schiller, Friedrich (32 - 7 - 0.683 - 0.429)
645	Wagner, Richard (11 - 1 - 0.702 - 0.0)
646	Aesop (22 - 4 - 0.714 - 0.5)
647	Sudermann, Hermann (14 - 1 - 0.715 - 0.0)
648	Milne, A. A. (Alan Alexander) (11 - 2 - 0.733 - 0.5)
649	Maugham, W. Somerset (William Somerset) (26 - 5 - 0.853 - 0.6)
650	Honig, Winfried (11 - 2 - 1.081 - 0.0)

Table A.2. *The genres that we use in our study.*

	Genre (No. documents - test size - consistency - accuracy)	Genre (No. documents - test size - consistency - accuracy)
0	World War II (11 - 6 - 0.08 - 0.667)	World War I (57 - 17 - 0.294 - 0.235)
1	Crime Fiction (27 - 7 - 0.125 - 0.857)	Art (14 - 1 - 0.299 - 1.0)
2	Historical Fiction (263 - 48 - 0.152 - 0.833)	Animal (16 - 2 - 0.313 - 1.0)
3	Western (76 - 18 - 0.153 - 0.611)	Children's Literature (158 - 26 - 0.33 - 0.462)
4	Horror (16 - 2 - 0.159 - 0.0)	Classical Antiquity (13 - 3 - 0.344 - 0.0)
5	Children's Book Series (354 - 75 - 0.176 - 0.853)	US Civil War (78 - 13 - 0.345 - 0.462)
6	Adventure (37 - 9 - 0.179 - 0.333)	Christmas (44 - 8 - 0.37 - 0.125)
7	Children's Fiction (269 - 58 - 0.188 - 0.879)	Fantasy (48 - 8 - 0.375 - 0.375)
8	Crime Nonfiction (20 - 5 - 0.188 - 0.0)	Poetry (22 - 4 - 0.391 - 0.0)
9	Science Fiction (447 - 87 - 0.211 - 0.851)	Travel (16 - 2 - 0.396 - 0.0)
10	Movie Books (37 - 6 - 0.221 - 0.0)	Children's Picture Books (35 - 8 - 0.424 - 0.5)
11	Biology (15 - 3 - 0.224 - 0.667)	Children's Instructional Books (12 - 2 - 0.44 - 0.0)
12	Children's History (23 - 4 - 0.224 - 0.25)	Harvard Classics (40 - 10 - 0.474 - 0.0)
13	Humor (82 - 19 - 0.225 - 0.474)	Best Books Ever Listings (54 - 10 - 0.514 - 0.3)
14	Precursors of Science Fiction (12 - 1 - 0.264 - 0.0)	One Act Plays (28 - 7 - 0.516 - 0.571)
15	School Stories (33 - 5 - 0.269 - 0.2)	Philosophy (56 - 9 - 0.541 - 0.778)

https://doi.org/10.1017/S0956792520000157 Published online by Cambridge University Press