The Character in the Letter:
Epistolary Attribution in
Samuel Richardson's *Clarissa*

Lisa Pearl, Kristine Lu, and Anousheh Haghighi
Department of Cognitive Sciences
3151 Social Science Plaza
University of California, Irvine
Irvine, CA 92697
Corresponding author email: lpearl@uci.edu

Abstract

Deliberate differences in how authors represent characters has been a core area of literary investigation since the dawn of literary theory. Here, we focus on epistolary literature, where authors consciously attempt to create different character styles through series of documents like letters. Previous studies suggest that the linguistic gestalt of an author's style – the author's *writeprint* – can be extracted from the various characters of an epistolary novel, but it is unclear whether individual characters themselves also have distinct writeprints. We examine Samuel Richardson's *Clarissa*, lauded as a watershed example of the epistolary novel, using a recently developed and highly successful authorship attribution technique to determine (i) whether Richardson can construct distinct character writeprints, and (ii) if so, which linguistic features he manipulated to do so. We find that while there are not as many distinct character writeprints as characters, Richardson does appear to have signature features he alters to create distinct character styles – and few of these features are the function word or abstract syntactic features typically comprising author writeprints. We discuss implications for other questions about character identity in *Clarissa* and character writeprint analysis more generally.

1 Introduction

Since the dawn of literary theory, deliberate differences in how authors represent characters has been a core area of investigation (Aristotle, 350 BCE). More recently, technological advances and interdisciplinary collaborations have expanded the methodologies used for literary scholarship, enabling investigations into more complex topics of authorship (e.g. The Federalist Papers: Adair, 1944; Mosteller and Wallace, 1963; Rokeach et al., 1970; Holmes and Forsyth, 1995; Tweedie et al., 1996; Bosch and Smith, 1998; Fung, 2003; Collins et al., 2004; Oakes, 2004; Rudman, 2005) and character psychology (Zunshine, 2010; Vermeule, 2011).

Computational stylistics (Milic, 1966; Stamatatos et al., 2001) has been a favored tool in literary scholarship for investigating authorial differences for at least two reasons: (i) its emphasis on character, and (ii) its ability to provide quantitative analysis without displacing the critic's powers of interpretation (Smith, 1989). The related field of authorship attribution often uses similar techniques based on author differences to determine an author's identity. This has been useful for contentiously attributed documents like the Federalist Papers, as well as in cases of authorship deception, where one writer attempts to consciously imitate another author's style (sometimes called *imitation attacks*: Brennan and Greenstadt, 2009; Pearl and Steyvers, 2012).

Here, we focus on epistolary literature, a genre defined by heightened mimetic qualities such as the lack of an omniscient narrator and the collection of "found documents" by each of the novel's characters. These qualities make it an intriguing domain for questions of character identity, style, and authorship. In general, authors are assumed to have a *writeprint* where the gestalt of their linguistic feature usage is distinctive (Li et al., 2006; Abbasi and Chen, 2008; Iqbal et al., 2008, 2010; Pearl and Steyvers, 2012). Even in epistolary novels where authors consciously attempt to create different character styles, the author's writeprint can be extracted. For example, Burrows (2005) successfully applied his prominent Delta technique to identify differences in the writing of Samuel Richardson's epistolary *Pamela* and Samuel Fielding's subsequent parody *Shamela*.

However, a related question concerns the characters within the epistolary novel itself: since they are all written by the same author, are they in fact distinct? That is, do different characters created by the same author have distinct character writeprints? While an author's writeprint may imprint each of the created characters, those characters could still be quite different stylistically; on the other hand, it could be that the author was unable to deviate significantly from his author writeprint for any of the characters. Notably, the linguistic features that comprise an author's writeprint have often been drawn from function words and abstract syntactic structures, because these are believed not to be consciously manipulable (Mosteller and Wallace, 1964; Burrows, 1987; Holmes et al., 2001; Binongo, 2003; Burrows, 2003; Juola and Baayen, 2005; Zhao and Zobel, 2005; García and Martin, 2007; Stamatatos, 2009; Lučić and Blake, 2015; Kestemont et al., 2015). This is why they become part of the linguistic signature of that author. So, we might expect that an author is unable to alter them even when writing as different characters. Instead, perhaps other linguistic features are altered, or perhaps the author is not truly able to distinguish characters stylistically at all.

Recently, van Dalen-Oskam (2014) examined precisely this question of distinctive character writeprints within the epistolary novels of famous Dutch women writers. While successful in iden-

tifying the author writeprints using existing techniques, van Dalen-Oskam was less satisfied with the results of the character writeprint analysis using existing tools like Bootstrap Consensus Trees (Eder, 2010; Eder and Rybicki, 2011) and Burrows's Zeta (Burrows, 2007). She concluded that convincing and objective computational methods do not yet exist for the task of identifying exact differences between character writers, and that "a lot of work will have to be done to find a (combination of) method(s) that will lead to verifiable and repeatable results," particularly beyond "words and their frequencies" (van Dalen-Oskam, 2014). One contribution of the current study is the application of a different writeprint technique used by Pearl and Steyvers (2012) for authorship deception, with what we feel are more satisfying results for issues surrounding character writeprints.

We focus our investigation on Samuel Richardson's *Clarissa*, one of the longest novels in English history, at over 1500 pages and nearly a million words. It is a novel rich with critical history and has been lauded as a watershed example of the epistolary novel, receiving a pedagogical renaissance for its expanded set of character authors (over 30) and insight into psychological realism (Zunshine, 2010). With *Clarissa*, Richardson sealed his reputation not only as a masterful writer cited for his editorial prowess (Price, 2000, p.27) but also as an ambitious publisher of literary insight. The novel centers around the beautiful and virtuous Clarissa Harlowe, a young lady caught between her ambitious, greedy family and the wiles of the dashing libertine Robert Lovelace (referenced mainly by his last name). Lovelace's consuming desire to possess Clarissa leads to increasingly dastardly and involved ploys of abduction and seduction. To illustrate the moral, familial, and psychological torment facing Clarissa as she stalwartly clings to her virtue, Richardson creates a diversity of meticulous epistles ranging from letters to legal documents to musical compositions to torn remnants of Clarissa's improvisational poetry. As noted in Richardson's postscript to *Clarissa*, the goal of this heterogeneous collection was to capture the "interesting personalities" of the characters represented while also allowing them to be "various", "natural", and "well distinguished". That is, a primary goal for Richardson was to make the characters distinct enough to realistically be separate people conversing with each other.

Given this rich dataset of character styles, we explore two basic questions about character writeprints in *Clarissa*. First, can Richardson construct distinct character writeprints at all? If so, it is useful to determine how distinct they are and whether there are as many character writeprints as there are characters. Second, if any distinct character writeprints were in fact created, which linguistic features did Richardson manipulate in order to do so?

We begin by describing the *Clarissa* epistolary corpus in more detail. We then briefly review the writeprint analysis method from Pearl and Steyvers (2012) (henceforth **PS**) that we apply, comparing it to other commonly used authorship techniques and highlighting its utility for automatically determining linguistic features indicative of particular characters. We next discuss the set of potential linguistic features that can comprise a character's writeprint, which the PS technique draws from to automatically construct a given character's writeprint. Our results suggest that Richardson is somewhat successful at creating distinctive character styles. However, there are not as many character writeprints as there are characters, and even the character writeprints discovered are not as distinct as typical author writeprints identified by the PS method. Nonetheless, there do appear to be signature features that Richardson tends to alter to create distinct character styles. In-

terestingly, few are the function word features that have traditionally comprised author writeprints (Mosteller and Wallace, 1964; Burrows, 1987; Holmes et al., 2001; Binongo, 2003; Burrows, 2003; Juola and Baayen, 2005; Zhao and Zobel, 2005; García and Martin, 2007; Kestemont et al., 2015) or the deeper syntactic features more recently gaining prominence (Stamatatos, 2009; Lučić and Blake, 2015). We discuss implications for other questions about character identity in *Clarissa* and character writeprint analysis more generally, concluding with suggestions for fruitful future work in this area.

2   Corpus: Richardson's *Clarissa*

The version of Richardson's *Clarissa* used in our research was downloaded from The Oxford Text Archive at *http://ota.ahds.ac.uk*, and contains 789 letters, comprising 918,624 words total.[1] There are samples from 35 distinct characters in total, including an epilogue by Richardson himself. Though nearly all samples are letters by a single author, several are not so neatly classified: (i) letters where it is undecided who wrote them, (ii) letters jointly written by two or more characters, (iii) letters written by one character pretending to be another, and (iv) a conclusion "supposedly written by" one of the characters, John Belford. Table 1 summarizes the distribution of letters across characters. It is clear that the majority of the letters (619 of 789, which is 78%) come from just a few of the 35 characters: the two central characters Clarissa Harlowe and Lovelace, and their respective confidantes, Anna Howe and John Belford, as highlighted in Fig. 1. Fig. 2 shows the distribution of words per letter for these four characters, which can range from tens of words to thousands of words.
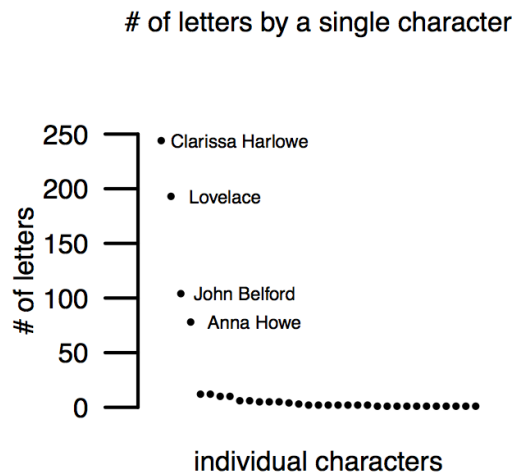

Figure 1: The number of letters each character wrote is shown, with the points representing the four main characters indicated. The remaining points correspond to the other characters who authored letters in *Clarissa*.

5

Table 1: Distribution of letters by character in *Clarissa*, indicating character name and the number of letters written by that character.

| Single Characters	#	Others	#
Clarissa Harlowe	244	Undecided	59
Lovelace	193	Two authors	4
John Belford	104	Lovelace as Clarissa	2
Anna Howe	78	Lovelace as Anna	1
Judith Norton, William Morden	12	Supposedly written by Belford	1
Arabella Harlowe, James Harlowe, Jr.	10		
Elizabeth Lawrance, John Harlowe	6		
Charlotte Harlowe, Charlotte Montague, Lord M	5		
Antony Harlowe	4		
Anabella Howe	3		
Antony Tomlinson, C. H. Hickman, Clarissa's Father, Dorothy Hervey, Elias Brand, Joseph Leman, R.D. Mowbray	2		
Alexander Wyverly, Arthur Lewen, Clarissa's Grandfather, Dolly Hervey, F. J. de la Tour, Hannah Burton, Patrick McDonald, Roger Solmes, Samuel Richardson, Tho Doleman, William Summers	1		

Character letter distributions
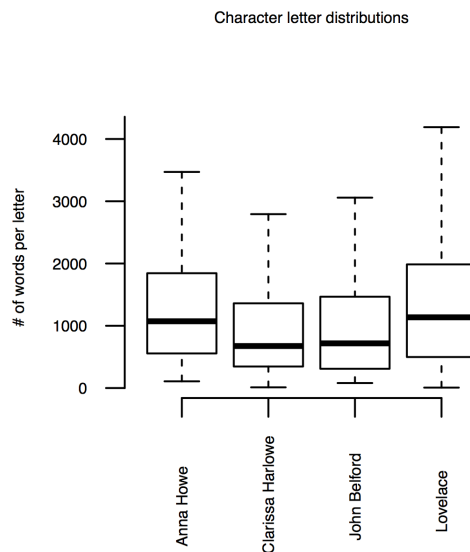
Figure 2: The distribution of words per letter for the four main characters.

3   Methods

3.1   Overview of the PS Authorship Method

We adopt the method used by Pearl and Steyvers (2012), a highly successful authorship approach that incorporates several aspects useful for character writeprint analysis, which are listed below in (1). The PS method is notable for using all these components together, though other authorship methods typically use some subset of them.

(1)   PS authorship components
  a.   Preprocessing the raw linguistic feature values to increase the perceived importance of those that are distinct for a given author
  b.   Using a variety of linguistic feature types
  c.   Utilizing a subset of the available features to form the writeprint
  d.   Allowing some writeprint features to matter more than others when making authorship decisions

We briefly review each PS method component in turn before describing the method in more detail.

3.1.1   Preprocessing Feature Values

The preprocessing step is somewhat similar to using the Kullback-Leibler Divergence (KLD), often called *relative entropy* (Zhao et al., 2006; Savoy, 2013, 2015), as well as to Burrows's Delta (Burrows, 2002; Hoover, 2004; Burrows, 2005, 2007; Argamon, 2008; Stamatatos, 2009; Savoy, 2013; Kestemont et al., 2015; Savoy, 2015) and the chi-squared method (Grieve, 2007; Savoy, 2015). These approaches compare the probability of a given feature value for the author in question against the probability of that feature value in a specified comparison set. For the PS method, the comparison set is the entire set of authors collectively; in contrast, for the KLD, Burrows's Delta, and chi-squared methods, the comparison set is a single author at a time. To our knowledge, no other authorship method currently uses this component implemented this way, though the PS implementation of it could be incorporated into any method that involves running an algorithm over feature values.

3.1.2   Linguistic Feature Types

Unlike methods that rely on word frequency alone (e.g. Burrows's Delta), the PS method allows linguistic features to range across a variety of character-level, word-level, syntactic, semantic, and formatting features (see Tables 7-17), similar to some previous approaches (see Stamatatos (2009) for a review, and Luyckx and Daelemans (2011) and Eder (2013), among others). Like the preprocessing component, this could be used with any method that involves running an algorithm over feature values, though several recent implementations have not done so (e.g. the KLD implementation of Zhao et al. (2006), the Nearest Shrunken Centroid (NSC) implementation of Jockers (2013), the Bootstrap Consensus Tree as applied by van Dalen-Oskam (2014), the Principle Component Analysis (PCA) implementation of Kestemont et al. (2015), and the KLD and chi-squared

implementations of Savoy (2015)).

3.1.3   Features in a Writeprint

The Sparse Multinomial Logistic Regression (SMLR) algorithm of Krishnapuram et al. (2005) used by the PS method has the ability to automatically determine which subset of the available features is most useful for making authorship decisions, like several other approaches (e.g. Burrows's Delta, NSC, PCA, and Support Vector Machines (SVMs)). This contrasts with methods like K-nearest neighbors (KNN) that obligatorily use the entire feature set. Importantly, the feature subset is what comprises the writeprint of any particular author. One notable ability of the SMLR algorithm is that it can potentially identify a different subset of features for each author writeprint, distinguishing it from methods that require all writeprints to use the same subset of linguistic features (e.g. PCA, NSC).

3.1.4   Using Writeprint Features to Make Authorship Decisions

Like some other approaches (e.g. SVM, PCA), the SMLR algorithm used by the PS method allows some writeprint features to matter more than others when making authorship decisions. A feature's importance to the algorithm is typically indicated by its weight. For example, a positive weight corresponds to a writeprint feature that is indicative of the author, with higher weights indicating features that are more useful for determining authorship.

3.2   Applying the PS Authorship Method for Character Writeprints

The basic representation of the problem the PS method will solve for character writeprints is whether a target letter (e.g. a letter by a character from *Clarissa*) is by the same character as the reference set of letters (e.g. letters by a single character from *Clarissa*). So, every data point will involve information derived from the sets of letters in (2), and the core decision is which character is the author.

(2)   Sets of documents used to create data points
  a.   *same* data point
       S = letter written by a character C1, Reference set R = all letters written by C1
  b.   *different* data point
       D = letter written by a character C2, Reference set R = all letters written by C1

For a "same author" data point (2a), the label should be the same as the character whose letters comprise the reference set (e.g. Clarissa Harlowe for the Clarissa Harlowe reference set). For a "different author" data point (2b), the label should be some other character – ideally the correct character, but at the very least not the character whose letters comprise the reference set (e.g. some non-Clarissa Harlowe character for the Clarissa Harlowe reference set).[2]

The SMLR classifier (Krishnapuram et al., 2005) is a supervised machine learning method that first trains on a collection of these data points that are labeled with the character author, learning

about the character writeprint for the character (C1) in the reference set. The classifier then attempts to apply this acquired writeprint knowledge to a new unlabeled collection of data points (the test set), which contains both same and different data points. If a character's writeprint is distinct, the classifier should perform well at identifying letters written by the character in the reference set (C1); in contrast, if a character's writeprint is not distinct, the classifier will perform poorly. A dataset is created for each character being investigated, with that character's letters used as the reference set R for all the data points in the set.

To create a data point, we follow the PS preprocessing procedure that increases the prominence of potentially distinctive feature values by determining if a particular feature value is unusual compared to the values characters typically have for that feature. We note that this procedure is applied to each feature separately. Specifically, we first calculate the probability of the feature value f_v occurring, given the distribution of feature values in the reference set of letters from the character in question: $p_{char} = p(f_v \mid distribution\ for\ character\ C1)$. Then, we compare p_{char} against the probability of that feature value occurring, given the distribution of feature values in the set of letters from all characters in *Clarissa*: $p_{all} = p(f_v \mid distribution\ for\ all\ characters)$.

This gives us a quantitative translation of how distinctive feature value f_v is for the character. If the feature value is unusual (e.g, f_1 in Fig. 3), the probability of f_v coming from the character in question will be higher than the probability of f_v coming from the entire population of characters (i.e. $p_{char} > p_{all}$). If the feature value is instead common to other characters as well, the probability of f_v coming from the character in question will not be higher (i.e. $p_{char} \approx p_{all}$). If the feature value is an aberration for this character (e.g, f_2 in Fig. 3), the probability of f_v coming from the character in question will be lower (i.e. $p_{char} < p_{all}$).

Each transformed feature value is calculated as follows, with the major aspects highlighted in Fig. 3. First, we log transform the feature values from the reference set of letters for that character (i.e. new value = log(raw value)), which creates a distribution of values that is roughly normally distributed, provided sample size is large enough. We then estimate the best-fitting normal distribution for this observed distribution that represents the feature distribution for this character's letters (CHAR normal distribution). We note that this is where the size of the letter set authored by a particular character ceases to matter (e.g., one character writing 244 letters while another character only writes 78 letters) – the only information extracted is the parameter values of the best-fitting normal distribution of the feature values in that characters letters. As long as a normal distribution can be reasonably estimated, the exact size of the character letter data set is irrelevant. We then apply the same process to the set of letters from all characters in *Clarissa*, generating the normal distribution ALL that represents that feature's distribution across all characters. We then calculate the probability that the observed feature value f_v in the target letter would be drawn from the CHAR distribution ($p(f_v|CHAR) = p_{char}$) and compare that against the probability that f_v would be drawn from the ALL distribution ($p(f_v|ALL) = p_{all}$) using the log-odds ratio in (3):

(3) $\log(\frac{p_{char}}{p_{all}}) = \log(\frac{p(f_v|CHAR)}{p(f_v|ALL)})$

A positive value means that p_{char} is larger than p_{all}, and so this feature value is unusual for the population of characters as a whole, but more typical for the character whose letters comprise the reference set. That is, it is more likely to be a distinctive feature for this character. A negative value

means that p_{char} is smaller than p_{all}, and so this feature value is more typical for the population of characters as a whole and not for the reference set character. That is, it is likely not to be a distinctive feature for this character. Fig. 3 demonstrates this for values f_1 and f_2, where f_1 would have a positive log-odds ratio while f_2 would have a negative one.
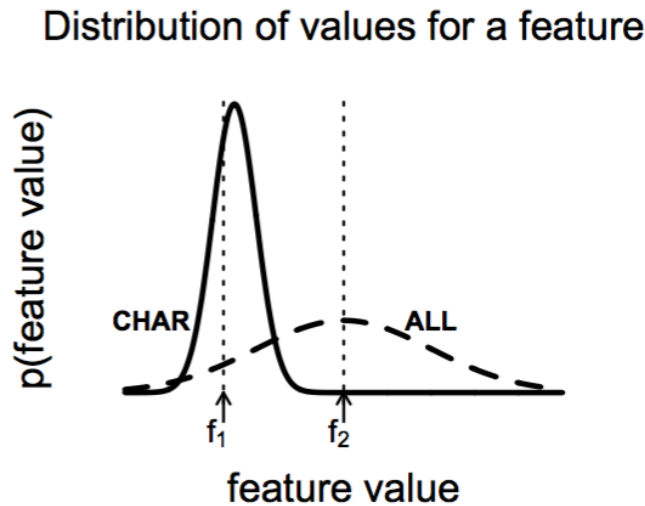
Distribution of values for a feature

Figure 3: Sample normal distributions derived from the log-transformed values for a single feature in the letters of a single character (CHAR) and the letters of all characters (ALL). Sample feature value f_1 is typical of CHAR but atypical of ALL, and so would have a positive log-odds ratio. Sample value f_2 is atypical of CHAR but typical of ALL, and so would have a negative log-odds ratio.

Because this preprocessing procedure is applied to each feature separately, the effective diagnosticity of each feature is assessed separately. Importantly, what matters is not how large or small the raw feature value is for a given feature, but how unusual it is compared to other feature values that occur.

So, a data point in a character's data set, derived from a target letter and a reference set of letters from that character, has the form in (4): a label indicating the character author of the target letter followed by a set of log-odds transformed feature values from that target letter, given the reference set.

(4)   Sample data points derived from target letters and a reference set of letters by Anna Howe
   a.   *Anna Howe*, 0.668627018303141, 3.12335486642762, 0.141156015976692, 0.336583176410039, ...1.82762529986441
   b.   *Clarissa Harlowe*, 0.666158689800687, 2.58610541486525, 0.141156015976692, 0.336898797155784, ... 1.21406973468136

A dataset consists of 789 data points, with every letter in *Clarissa* used as a target letter. So, some portion of a dataset will be labeled with the reference author (e.g. Anna Howe in (4)), while the rest

will be labeled with other characters. Because the preprocessing procedure requires the reference set to be of some size in order to more accurately estimate a normal distribution for an individual character's reference set of letters (CHAR), we restricted our analysis to characters with fifty or more letters in *Clarissa*. This confined our analysis to the four characters with the most letters and so yielded four distinct datasets: Clarissa Harlowe, Lovelace, John Belford, and Anna Howe. The SMLR classifier was then run on each constructed dataset.

We note that the SMLR classifier requires a parameter λ that controls how strongly the classifier is biased to select a subset of the available linguistic features for a character's writeprint. Larger values of λ lead to writeprints consisting of fewer features, while smaller values lead to writeprints consisting of more features. A value of 0.1, which is a fairly strong bias to prefer fewer features, led to the best performance in a pilot analysis, and so we used this for our analyses below.

We used the SMLR java implementation available from *http://www.cs.duke.edu/~amink/software/ smlr/*, implementing ten-fold cross-validation to evaluate the classifier's performance. In ten-fold cross-validation, the dataset is divided into ten "folds" of equal size, with each fold containing approximately the same distribution of labeled data as the entire data set. So, for example, since 244 of 789 letters are Clarissa Harlowe letters, each fold contained 78 or 79 target letters and approximately 24 target letters in each fold were Clarissa Harlowe letters. The classifier then makes ten passes through the dataset, training on the data in nine of the ten folds and testing on the data in the remaining fold, with each fold taking a turn as the test fold. So, for every pass, the classifier learns what it can from $\frac{9}{10}$ of the dataset about how the character labels are determined based on the preprocessed feature values, and then attempts to label the data points in the remaining $\frac{1}{10}$ with the appropriate character. One benefit of cross-fold validation is that results are obtained for every single data point in the data set, as each data point will be in one of the test folds. This guards against results being impacted by a particularly easy or difficult test set, since ten test sets are used, and the average of the ten test sets is the final score.

3.3   Linguistic Features Available for Character Writeprints

The PS method uses the SMLR classifier to automatically construct writeprints from a set of available linguistic features and bases its authorship decisions on those writeprints. We extracted a set of 122 linguistic features of seven different kinds, shown in Tables 7-17 in the Appendix: 22 character-level features, 5 word-level features, 28 syntactic category features that could be viewed as contentful (e.g. nouns), 24 syntactic category features that could be viewed as functional (e.g. prepositions), 32 syntactic structure features (e.g. passives), 4 formatting features (e.g. italicized words), and 7 semantic features (e.g. endearments). Similar to Pearl and Steyvers (2012), these are all stylometric features that can be extracted automatically using freely available natural language processing software[3] and scripts written in a text manipulation programming language like PERL. Notably, many of these features can be easily and automatically extracted from languages besides English, provided the natural language processing software is available in the desired language (i.e., the character-level features, the word-level features, several content syntactic category features, several syntactic structure features, and the formatting features). However, we do note that many of the remaining features were manually identified using knowledge of English grammar (some content syntactic category features, several functional syntactic category features, and

some syntactic structure features). Additionally, the semantic features were manually identified using domain-specific knowledge of *Clarissa*, rather than using the topic-modeling approach of Pearl and Steyvers (2012), due to the relatively small size of the corpus. In general, when applying the PS method to character writeprints, we believe it is likely more expedient for literary scholars to draw on their own knowledge of the particular literature under investigation when identifying potentially useful semantic features. For scholars studying works in other languages, this would require manual identification of the relevant semantic features in those languages.

4   Results

We present two kinds of results: (i) quantitative results relating to character writeprints, based on applying the PS method, and (ii) qualitative interpretations of those results that relate to character distinctiveness and signature writeprint features manipulated by Richardson.

4.1   Character Distinctiveness

For each of the four characters who wrote over fifty letters in *Clarissa*, we can examine how distinctive that character's style is by assessing the ability of the PS method to correctly label a target letter with the correct character, given a reference set of letters by that character (e.g. label a letter by Clarissa Harlowe as being by her, when referenced against a set of letters by her). Notably, the PS method had an average success rate of 100% on one data set and 89% on another in Pearl and Steyvers (2012)'s study that attempted to identify different authors, so it is very accurate when there are in fact different authors writing. If the character writeprints in *Clarissa* are as distinct as author writeprints typically are, we would expect similar performance.

From the results of the SMLR classifier for each character, we can derive a confusion matrix as in Table 2, where the rows represent the true character writer of the letter and the columns represent the character writer labeled by the SMLR classifier. The four quantities correspond to the distinctions in signal detection theory, representing (i) *true positives* (a): the number of letters where the SMLR correctly labeled the letter by the target character as being by the target character, (ii) *false negatives* (b): the number of letters where the SMLR incorrectly labeled a letter by the target character as being by a different character, (iii) *false positives* (c): the number of letters where the SMLR incorrectly labeled a letter not by the target character as being by the target character, and (iv) *true negatives* (d): the number of letters where the SMLR correctly labeled a letter not by the target character as not being by the target character.

Standard metrics used in computational linguistics to gauge the success of a classifier are precision and recall, defined as in (5a) and (5b), and combined into a single summary statistic known as the F-score using the harmonic mean definition in (5c). Precision, recall, and F-score all range between 0 and 1, with 1 being perfect performance. The intuitive interpretation of precision is how accurate identification is, while the intuitive interpretation of recall is how complete identification is. Ideally, a classifier will be both very accurate and very complete in its identification, yielding a high F-score.

Table 2: The performance of the SMLR classifier can be summarized using a confusion matrix, where the rows represent the true identity of the target letter's author and the columns represent the SMLR-labeled identity of the target letter's author. Correct labels are a (true positives) and d (true negatives), while incorrect labels are b (false negatives) and c (false positives).

		Labeled	
		character	non-character
True	character	a	b
	non-character	c	d

(5)   Evaluation metrics, with quantities from Table 2

   a.   Precision = $\frac{\#\ correctly\ labeled\ as\ character\ (a)}{\#\ labeled\ as\ character\ (a+c)}$:

   b.   Recall = $\frac{\#\ correctly\ labeled\ as\ character\ (a)}{\#\ should\ be\ labeled\ as\ character\ (a+b)}$:

   c.   F-Score = $2 * \frac{precision * recall}{precision + recall}$

Table 3 shows the F-scores for each of the four characters investigated, while the detailed information from the confusion matrices as well as the precision and recall scores for each character appear in Appendix B. Most notably, *none* of the characters seem to have an easily identifiable style – the highest performance by F-score is 0.558. This suggests that it is not easy for a single author, such as Richardson, to differentiate the writing styles of different characters. Still, there seem to be natural classes of characters: (i) those that are more distinctive (Lovelace: 0.558), Clarissa Harlowe: 0.535), (ii) those that are somewhat distinctive (John Belford: 0.396), and (iii) those that are not very distinctive at all (Anna Howe: 0.217).

There are at least two reasons why a character's writeprint may not be very distinctive. First, a character may simply be a hodgepodge of multiple characters' styles, and so would overlap with all other character styles. Alternatively, a character may be a derivative of other character styles and have a writeprint that borrows from the main writeprint of those other character styles.

We can discern between these options by looking at the common confusions that emerge from the SMLR confusion matrix, shown in Table 3. Any character which the SMLR confused the target character with more than 10% of the time (either for the precision calculation, the recall calculation, or both) is shown. We observe that both John Belford and Anna Howe are confused with Clarissa Harlowe and Lovelace, but not with each other; we interpret this to mean that John Belford's and Anna Howe's writeprints are distinct from each other. Moreover, while Lovelace is confused with John Belford, he is not confused with Anna Howe; we interpret this to mean John Belford's writeprint comes primarily from Lovelace's.[4] Similarly, while Clarissa Harlowe is confused with Anna Howe, she is not confused with John Belford; we interpret this to mean Anna Howes writeprint comes primarily from Clarissa Harlowe's.

Given the quantitative information from the F-scores and qualitative information from the systematic confusions, we suggest that Richardson was able to generate two main styles, one for each of the two central characters (Lovelace, Clarissa Harlowe). He then derived additional charac-

ter styles from those two main styles (John Belford from Lovelace, Anna Howe from Clarissa Harlowe). Narratively, this makes intuitive sense because Belford and Anna serve primarily as "sounding walls" for Lovelace and Clarissa, respectively, as the moral conflict in the novel escalates.

Because of this, an interesting subsequent analysis is to calculate the F-scores for the two main characters while allowing the letters derived from their styles to count as instances of the main character letters – that is, John Belford's letters counted as instances of Lovelace's letters and Anna Howe's letters counted as instances of Clarissa Harlowe's letters. This increases the F-scores of both main characters: Lovelace's F-score increases to 0.586 while Clarissa Harlowe's F-score increases to 0.575. Though this improvement is non-trivial, these character F-scores are still far below what we see when we compare the writings of two different authors (where F-scores are at 0.890 and above, based on Pearl and Steyvers (2012)). A possible reason for this involves the linguistic features that Richardson is able to manipulate to create different character styles. We examine these next.

4.2 Signature Features

For each character (whether distinctive or not), we can examine the most distinctive features according to the SMLR analysis, since the classifier learned not only which features comprised a character's writeprint but also how important each of those features was to the writeprint. The importance of a given feature is indicated by the weight learned for it; the features with the highest positive weight are the ones that influenced the SMLR's decision the most when deciding to label a target letter based on a particular character's reference set. Though the SMLR bases its decision on all non-zero weighted features, these top features can serve qualitatively as the "signature" writeprint features for that character's style, as they matter the most to the classifier's decision. From this, we can assess how similar signature features are across character styles. Table 3 lists the signature features for each character style, indicating which ones are shared across character styles and which specific characters a given signature feature is associated with when more than one character utilizes that signature feature.

There do seem to be common features Richardson prefers to manipulate, and every character (distinctive or not) has several. The most commonly manipulated linguistic features are shown in Table 4. Interestingly, few of these commonly manipulated features include what would be considered function words, which have often been a core distinguishing feature of author writeprints (Mosteller and Wallace, 1964; Burrows, 1987; Holmes et al., 2001; Binongo, 2003; Burrows, 2003; Juola and Baayen, 2005; Zhao and Zobel, 2005; García and Martin, 2007; Stamatatos, 2009; Lučić and Blake, 2015; Kestemont et al., 2015). Only one feature (frequency of all function words together) is of this kind.

Similarly, few are the abstract syntactic features that have also been central to more recent authorship studies (Lučić and Blake, 2015). For example, while the average lengths of different syntactic phrases were potential writeprint features that would capture nuances of verbosity specific to those syntactic structures, none of these features were identified as common writeprint features that Richardson manipulated. The three he manipulated which are closest to this kind of feature are the frequency of clauses, the frequency of fragments, and the frequency of wh-adverb phrases.

14

Table 3: Distinctiveness of main characters, in descending order by F-score, which is a summary statistic of classifier performance. Signature features for each character (SMLR weight ≥ 1.4) are also listed in descending order of strength, according to the writeprints identified by the SMLR classifier. Signature features appearing for more than one character are **bolded**, with the initials of the character(s) that share(s) that signature feature in parentheses. Common confusions (accounting for ≥10% errors for precision and/or recall) are also shown.

Character	F-score	Signature features	Common conf
Lovelace (L)	0.558	**verb frequency (CH, AH)**, alphabetic character frequency, verb phrase frequency, **function word frequency (AH)**, noun phrase frequency, **fragment frequency (JB)**, total characters, **punctuation frequency (CH)**, total words, **wh-adverb phrase frequency (JB)**, **title frequency (JB)**, **word length (JB)**, familial term frequency, gerund or present participle frequency, infinitive to frequency	CH, JB
Clarissa Harlowe (CH)	0.535	**verb frequency (L, AH)**, **clause frequency (JB)**, **punctuation frequency (L)**, first person pronoun frequency, universal determiner frequency, contraction frequency	L, AH
John Belford (JB)	0.396	**clause frequency (CH)**, parenthesis frequency, noun frequency, colon frequency, **fragment frequency (L)**, **title frequency (L)**, **endearment frequency (AH)**, adverb phrase length, **word length (L)**, **wh-adverb phrase frequency (L)**	CH, L
Anna Howe (AH)	0.217	**function word frequency (L)**, **endearment frequency (JB)**, **verb frequency (L, CH)**, wh-noun phrase frequency, second person pronoun frequency, emdash frequency, noun phrase frequency	CH, L

Interestingly, two of these three syntactic structure features have cues that are somewhat contentful. Sentence fragments represent an incomplete structure, and an incomplete syntactic structure leads to an incomplete sentence meaning. Likewise, wh-adverb phrases are headed by *when*, *where*, *why*, and *how*, and these words are fairly easy to define (e.g., the time something happened = *when*). So, both sentence fragments and wh-adverb phrases may be structural features that are easier to consciously recognize and manipulate. Clause frequency may operate this way as well, since clause frequency within a sentence can be increased by either conjoining main clauses together (e.g. *I like this* and *I want to read more*) or embedding clauses in main clauses (e.g. I think *that I like this*). In either case, a complete thought (represented by the additional clause) is added. Thus, clause frequency may also be straightforward to consciously manipulate.

The remaining commonly manipulated features range over syntactic categories based on con-

tent words (frequency of verbs), semantic categories (frequency of endearments and titles), and character-level features (frequency of punctuation, average word length). These may also be easier to consciously recognize and manipulate.

This lends support to the idea that both functional category and abstract syntactic structure usage are more unconscious and therefore more indicative of an author's genuine identity. If functional category and abstract syntactic structure usage are harder to consciously manipulate, and these aspects are often at the heart of an author's writeprint, this could be one reason why character writeprints aren't as distinctive as author writeprints typically are – even for an expert like Richardson. Instead, Richardson focused on features he could consciously manipulate, which are other feature types.

Table 4: Linguistic features most commonly manipulated by Richardson, including the type of feature, number of characters where the linguistic feature is among the set of signature features, and the specific characters whose writeprints contain the signature feature. Characters with distinctive writeprints are indicated with an asterisk (*).

Feature	Type	# char	Characters
verb frequency	syntactic (content)	3	*Lovelace, *Clarissa Harlowe, Anna Howe
clause frequency	syntactic (structure)	2	*Clarissa Harlowe, John Belford
fragment frequency	syntactic (structure)	2	*Lovelace, John Belford
wh-adverb phrase frequency	syntactic (structure)	2	*Lovelace, John Belford
function word frequency	syntactic (function)	2	*Lovelace, Anna Howe
endearment frequency	semantic	2	John Belford, Anna Howe
title frequency	semantic	2	*Lovelace, John Belford
punctuation frequency	character (punctuation)	2	*Lovelace, *Clarissa Harlowe
word length	character (word)	2	*Lovelace, John Belford

Next, *how* did Richardson manipulate the signature features for each character? For each signature feature in a character's writeprint, we can examine how the distribution of that feature's values in a character's letters compares to the distribution of that feature's values in the entire set of character letters. In particular, a simple analysis is whether that character's feature value was *generally* higher or lower than that feature's value in the character population, as measured by average or median feature value. If this occurs, it indicates one simple way that Richardson manipulated that feature to create an aspect of the character's writeprint. We note that not all signature features will have a distribution that shows up under this analysis because there are many

ways for a feature distribution to be distinctive, and having an average or median value that is higher or lower than the population average or median value is merely one of them.[5] Nonetheless, this is a convenient analysis to try as it lends itself well to verbally summarizing a character's style (e.g. one character uses more present participles and fewer titles), and generating content in that character's style (e.g. to emulate this character, use more present participles and fewer titles).

For each signature feature, we compared the character's median value against the population's median value, and the character's average value against the population's average value. If the character's median and/or average value was at least 10% higher or lower than the population's median and/or average value for that feature, we included it in the set of signature features in Table 5 that have an easily describable pattern of usage. For example, Lovelace's title frequency had a median value of approximately 0.004 and an average value of approximately 0.0000085. The population median value is approximately 0.006 while the population average value is approximately 0.000074. So, Lovelace's median value is less than the population median value by over 33% (0.004/0.006 - 1 = -0.333) and Lovelace's average value is less than the population's average value by over 88% (0.0000085/ 0.000074 - 1 = -0.885). This indicates that Lovelace's style uses titles less frequently than other characters, based on both the median and average values for this feature. Table 5 summarizes the results of this analysis for the four characters investigated.

Table 5: Signature features for each character that have an average or median characteristic usage that is either significantly higher (+10%) or lower (-10%) than the character population average or median. If average and median usage differ, [avg] or [median] indicate which one behaves which way. Distinct usage shared across character signature features is **bolded**, with the initials of the character(s) that share(s) that distinctive average/median usage in parentheses. If the shared distinctive usage is in the opposite direction (e.g. the first character's is + and the second character's is -), the character in parentheses is also *italicized*. Asterisks (*) indicate signature feature usage that is at least 100% higher/lower than the character population average and/or median.

Character	Signature features: Distinctive on average or median
Lovelace (L)	+: *gerund or present participle frequency, *infinitive *to* frequency, total characters, total words
	-: familial term frequency, **title frequency (JB [avg],** *JB [median]*)
Clarissa Harlowe (CH)	+: *contraction frequency, *first person pronoun frequency
	-: universal determiner frequency
John Belford (JB)	+: adverb phrase length, *colon frequency (CH), *parenthesis frequency, ***title frequency [median] (L)***
	-: ***endearment frequency (AH)***, **title frequency [avg] (L)**
Anna Howe (AH)	+: *emdash frequency, ***endearment frequency (JB)***, *second person pronoun frequency, *wh-noun phrase frequency
	-: None

From this analysis, we can observe two notable stylistic choices made by Richardson, both of

which appear to distinguish character styles. The first distinguishing feature is the use of endearments, as *endearment frequency* is a signature feature for both John Belford from Anna Howe, but in the opposite direction: Belford tends to use relatively fewer endearments while Anna tends to use relatively more. Interestingly, this is not a signature feature for either of the main writeprints (Lovelace and Clarissa), so it is unlikely to be something accidentally transferred from the main writeprints to the derived writeprints. Instead, it is more likely to be a conscious choice by Richardson to distinguish these two characters, who write more letters than any others except for Lovelace and Clarissa.

The second distinguishing feature is the use of titles (*title frequency*), which is manipulated for both Lovelace and John Belford. Lovelace always uses titles less frequently, while Belford's letters show a more nuanced pattern: his average usage is less frequent, but his median usage is more frequent. This suggests that Belford's letters typically use titles more frequently (i.e. many of Belford's letters have more titles than a typical character's letter), but there are a few outlier letters that use titles far less frequently than a typical character's letter. These outlier letters would lower the average title frequency while leaving the median title frequency relatively unaffected (see endnote 5 about the relationship between average and median usage). A potential interpretation of this pattern is that the usage of titles in letters is something Richardson felt was masculine, and so he consciously manipulated it when writing letters by male characters.

More generally, this analysis provides simple rubrics for how to write in the style of a specific character. For example, a message by Lovelace should contain few titles or familial terms, be verbose, and use both present participles and infinitival *to*. Table 6 provides sample messages that obey these rubrics, thereby representing "prototypical" examples of these characters' styles.

5   General Discussion

Using a state-of-the-art authorship classification approach (Pearl and Steyvers, 2012), we discovered that Richardson was able to create two distinct character writeprints for the four characters examined. This suggests that while it is possible for an author to make some distinctive characters writeprints, it is non-trivial to do so for each character.

Notably, the character writeprints Richardson is able to create are not as distinctive as author writeprints typically are. A related observation is that the features Richardson most often manipulates to create these character writeprints are not the functional or abstract syntactic features that have been prominent for authorship studies. We suggest that this may be due to the accessibility of these features. That is, the reason they are so often used for author writeprints is precisely because they are not easy to consciously manipulate. In contrast, when a single author is creating several character writeprints, the manipulated features may naturally be the ones that *are* consciously accessible.

5.1   *Applying the PS Approach for Related Authorship Questions in Clarissa*

In addition to these discoveries about the character writeprints in *Clarissa*, we can also answer interesting questions about literary deception in this particular epistolary novel. Notably, there

Table 6: Example messages from each character that follow the rubrics derived from each character's signature features.

Character	Message
Lovelace	And then, what a comely fight, all kneeling down together in one pew, according to eldership, as we have seen in effigie, a whole family upon some old monument, where the honest chevalier, in armour, is presented kneeling, with uplift hands, and half a dozen jolter-headed crop-eared boys behind him, ranged *gradatim*, or step-fashion, according to age and size, all in the same posture–Facing his pious dame, with a ruff about her neck, and as many whey-faced girls, all kneeling behind *her*: An altar between them, and an opened book upon it: Over their heads semilunary rays darting from gilded clouds, surrounding an atchievement-motto, IN COELO SALUS– or QUIES–perhaps, if they have happened to live the usual married life of brawl and contradiction. (*http://ota.ox. ac.uk/text/4363.html*)
Clarissa Harlowe	You'll observe, that altho' I have not demanded my estate in form, and of my trustees, yet that I have hinted at leave to retire to it. How joyfully would I keep my word, if they would accept of the offer I renew!–It was not proper, I believe you'll think, on many accounts, to own that I was carry'd off, against my inclination. (*http://ota.ox.ac.uk/text/ 4360.html*)
John Belford	He succeeds, takes private lodgings for her at Hackney; visits her by stealth, both of them tender of reputations, that were *extremely* tender, but which neither had quite given over; for rakes of either sex are always the last to condemn or cry down themselves: Visited by nobody, nor visiting: The life of a thief, or of a man beset by creditors, afraid to look out of his own house, or to be seen abroad with her. And thus went he on for twelve years, and, tho' he had a good estate, hardly making both ends meet; for, tho' no glare, there was no oeconomy; and besides, he had every year a child, and very fond of them was he. But none of them lived above three years: And being now, on the death of the dozenth, grown as dully sober, as if he had been a real husband, his good Mrs. Thomas (for he had not permitted her to take his own name) prevailed upon him, to think the loss of their children a judgment upon the parents for their wicked way of life... (*http://ota.ox.ac.uk/text/4361.html*)
Anna Howe	I HAVE both your letters at once. It is very unhappy, my dear, since your friends will have you marry, that such a merit as yours should be addressed by a succession of worthless creatures, who have nothing but their pre?umption for their excuse. That these presumers appear not in this very unworthy light to some of your friends, is, because their defects are not ?o striking to them, as to others.–And why? Shall I venture to tell you?–Because they are nearer their own standard.–*Modesty*, after all, perhaps has a concern in it; for how should they think, that a *niece* or a *sister* of *theirs* (I will not go higher, for fear of incurring your displeasure) should be an angel? (*http://ota.ox.ac.uk/text/4358.html*)

are three letters in which one character is pretending to be another: Lovelace writes one letter as Anna Howe and two letters as Clarissa Harlowe. This presents an intriguing layering effect for character writeprints: Richardson is attempting to create the character writeprint for a character

(Lovelace) who is attempting to imitate another character's writeprint (Anna or Clarissa). A very basic question is whether Richardson successfully shifted Lovelace's writeprint so that it appeared not to be Lovelace.

We applied the PS method to each of these three letters, using Lovelace's letters as the reference set. If Richardson was successful at altering Lovelace's writeprint, the SMLR classifier should *not* identify any of those letters as having been written by Lovelace. This was indeed the case for all three letters, meaning that Richardson effectively masked Lovelace's style for those letters.

Given this, was Lovelace successful in his deception as Anna and Clarissa? To answer this, we applied the PS method to these three letters, using Anna's letters as the reference set for the letter impersonating Anna and Clarissa's letters as the reference set for the letters impersonating Clarissa. Here, the deception seemed to fail. The letter impersonating Anna was not labeled as being by Anna when compared against the Anna reference set. Similarly, the letters impersonating Clarissa were not labeled as being by Clarissa when compared against the Clarissa reference set. So, Lovelace's deception was incomplete in this sense – though perhaps that was Richardson's intention. In particular, Richardson could have intended for the reader to recognize that the style wasn't quite the purported one in each case (Anna or Clarissa, respectively).

Additionally, there is a single letter that is "supposedly written" by John Belford. Yet, when compared against the Belford letters as a reference set, this letter was not labeled as being by Belford. This suggests that Richardson was unsuccessful in his portrayal of Belford as the writer for this letter, whether intentionally or unintentionally. Richardson may have intended for it to be written by one of the other characters; if so, it would have a writeprint matching one of these other characters. However, when compared against other character reference sets with ten or more letters (Anna Howe, Arabella Harlowe, Clarissa Harlowe, James Harlowe Jr, Judith Norton, Lovelace, William Morden), this letter was also not identified as being the writeprint of any of those characters. So, in general, this letter does not seem to be an effective portrayal of any easily identifiable character from *Clarissa*.

5.2 Applying the PS Approach More Generally for Character Writeprints

We believe the PS approach for discerning character writeprints in epistolary novels can be used to answer several questions of interest to literary scholars, including epistolary novel technique (both Richardson's and that of other epistolary novel authors), comparative evaluations of epistolary author skill, approaches to constructing character writeprints, and cues to author identity.

With respect to Samuel Richardson in particular, how skilled is he in his other epistolary novels (*Pamela, The History of Sir Charles Grandison*) at creating the appropriate number of character writeprints? Using the same PS approach, we can examine how distinct the character writeprints are and whether there are both main and derived writeprints, as in *Clarissa*. We can also investigate whether Richardson manipulates the same signature linguistic features as he did in *Clarissa*, and if these features tend to exclude the functional and abstract syntactic features common in author writeprints.

Additionally, we can apply the PS approach to investigate whether character writeprints are sensitive to major plot changes. This nuanced question would be particularly worthwhile to examine in *Clarissa*, as there is an abrupt letter marking the novel's astonishing climax when relationships

and alliances shift, particularly among the four central characters. Do the character writeprints reflect these changing alliances, e.g. character writeprint similarities aligning with current character alliances?

With respect to other writers of epistolary novels, especially Richardson's contemporaries like Jane Austen, Aphra Behn, Fanny Burney, James Howell, Frances Brooke, and Mary Shelley, how skilled are these other writers at creating character writeprints? Using the PS approach, we can determine how distinct these writeprints are, if there are the appropriate number, whether there are main and derived writeprints, which signature features are manipulated, and the nature of these signature features. We can also compare the signature features used by other epistolary authors to those we discovered for Richardson in *Clarissa*. If the same features or same types of features are commonly manipulated, this suggests that those are core features (or features types) for character writeprints within epistolary literature.

The PS approach also allows us to explore the impact of author identity in a unique way, based on the *types* of features that distinguish character writeprints and author writeprints. In epistolary novels, authors must subsume their own identity in order to generate a distinct set of character identities. Yet, our analysis of Richardson's *Clarissa* suggests that the features manipulated to create character writeprints differ from the features typically comprising author writeprints. Given this, it may be that an author consciously manipulates certain features to generate character writeprints while leaving the features that are not consciously accessible alone. In other words, the author's writeprint features may remain the same across all character writeprints generated by that author. Suggestive evidence for this view comes from Burrows (2005), who successfully distinguished Richardson's epistolary novel *Pamela* from Fielding's parody *Shamela*, and from van Dalen-Oskam (2014), who successfully distinguished the epistolary novels of two Dutch women writers. If the author writeprint features do in fact remain the same across character writeprints, it should be the case that character writeprints generated by one author are as distinct from character writeprints generated by another author as the author writeprint of the first author is from the author writeprint of the second author. This is a prediction that can be empirically tested with the quantitative approach we have used here, with intriguing implications for the cues to author identity if validated.

6   Conclusion

We have demonstrated how to apply current machine learning techniques to answer questions in literary scholarship related to authorship. This case study focuses on issues surrounding character authorship in the landmark epistolary novel *Clarissa*, yielding both quantitative and qualitative results about the ability of the innovative Samuel Richardson to develop distinct character styles within a novel. The particular machine learning approach we use can serve as a reliable tool for investigating other literary questions surrounding both the style of individual characters and the style of the author who creates those characters.

Notes

[1]This dataset is available at *http://www.socsci.uci.edu/~lpearl/CoLaLab/corpora/Richardsons_Clarissa.zip*, with letters organized by character.

[2]We note that we present results from an "n-way" classification task (where one of n characters is selected for each data point), but this task could also be set up as a simpler binary classification task where the goal is to label a letter as simply by the "same" author or by a "different" author than the reference set. Interestingly, we achieve better results with the harder n-way classification than the easier binary (2-way) classification, suggesting that it is useful to know which specific other character a letter was written by, rather than simply knowing that it was written by a different character from the the author of the reference set letters. This may be a specific instance of the more general situation where a seemingly harder problem is actually easier than a seemingly simpler problem because of the subtle information available in the data available for the harder problem (e.g. joint inference in cognitive development: Dillon et al., 2013; Doyle and Levy, 2013; Feldman et al., 2013; Börschinger and Johnson, 2014). In addition, the n-way classification task allows us to see which specific characters a given character is confused with, which is an important aspect of our character writeprint analysis.

[3] We used the Stanford Part-of-Speech Tagger (available at *http://nlp.stanford.edu/software/tagger.shtml*) to iden-tify syntactic categories and the Stanford Parser (available at *http://nlp.stanford.edu/software/lex-parser.shtml*) to identify syntactic structures.

The list of syntactic category tags from the Stanford Part-of-Speech tagger used is as follows, with an example of each tag in parentheses: CC = coordinating conjunction (***and***), CD = cardinal number (***one** penguin*), DT = determiner (***the***), EOS = end of sentence marker (*theres a penguin here**!***), EX = existential there (***there's** a penguin here*), FW = foreign word (***hola***), IN = preposition or subordinating conjunction (***after***), JJ = adjective (***good***), JJR = comparative adjective (***better***), JJS = superlative adjective (***best***), LS = list item marker (***one, two, three, ...***), MD = modal (***could***), NN = singular or mass noun (***penguin***, ***ice***), NNS = plural noun (***penguins***), NNP = proper noun (***Jack***), NNPS = plural proper noun (*There are two **Jacks**?*), PDT = predeterminer (***all** the penguins*), POS = possessive ending (*penguin**'s***), PRP = personal pronoun (***me***), PRP\$ = possessive pronoun (***my***), RB = adverb (***easily***), RBR = comparative adverb (***later***), RBS = superlative adverb (***most** easily*), RP = particle (*look it **up***), SYM = symbol (*this **=** that*), TO = to (*I want **to** go*), UH = interjection (***oh***), VB = base form of verb (*we should **go***), VBD = past tense verb (*we **went***), VBG = gerund or present participle (*we are **going***), VBN = past participle (*we should have **gone***), VBP = non-3rd person singular present tense verb (*you **go***), VBZ = 3rd singular present tense verb (*he **goes***), WDT = wh-determiner (***which** one*), WP = wh-pronoun (***who***), WP\$ = possessive wh-pronoun (***whose***), WRB = wh-adverb (***how***).

The list of phrase-structure tags the Stanford Parser used is as follows, with an example of each tag in parentheses: S = declarative sentence (***I like penguins***), SINV = sentences with subject-auxiliary inversion (***Never have I seen such penguins!***), SBAR = embedded clauses (*I like penguins **that are cute.***), INTJ = interjection (***um***), FRAG = fragment (***See penguins in the***) RRC = reduced relative clause (*penguins **not presently swimming***), SBARQ = wh-questions (***What did you see?***), SQ = yes/no questions (***Did you see that?***), ADJP = adjective phrase (***outrageously cute***), ADVP = adverb phrase (***rather sweetly***), CONJP = multi-word conjunctions (*...**as well as**...*), LST = list marker (***one, two, three***), NAC = not a constituent (***in the back of my mind it***), NP = noun phrase (***those penguins***), NX = sub-noun phrase (***other people and***), PP = preposition phrase (***with the penguins***), PRN = parenthetical (*Those penguins (***I really like them***)*), PRT = particle (*look it **up***), QP = quantifier phrase (***a little bit more***), UCP = unlike coordinate phrase (***from that, but thats why***), VP = verb phrase (*We **like penguins***), WHADJP = wh-adjective phrase (***How hot** is it?*), WHADVP = wh-adverb phrase (***How** are you?*), WHNP = wh-noun phrase (***Who** are you?*), WHPP = wh-preposition phrase (***With whom** did you see it?*), X = unknown phrase.

[4]This inference is also supported by noting that four of the six signature features of John Belford's writeprint that are shared by other characters are shared *only* by Lovelace's writeprint (fragment frequency, title frequency, word length, and wh-adverb phrase frequency). See discussion in the next section for how signature features were derived for each character.

[5]We note that while average and median population values generally align, sometimes they do not. This is because average values don't factor out the effect of outliers that shift the average value significantly up or down, while median values do. For example, suppose a set of 100 values has ten very low values around 0, while the remaining 90 values have an average of 50. The average of this set is about 45 while the median is likely to be around 50. A value of 46 is

then 2% higher than the average (46/45 - 1 = 0.02) but 8% lower than the median (46/50 - 1 = -0.08).

References

Abbasi, A. and Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Transactions on Information Systems (TOIS), 26(2): 7.

Adair, D. (1944). The authorship of the disputed Federalist Papers: Part II. The William and Mary Quarterly: Magazine of Early American History, Institutions, and Culture: 235–264.

Argamon, S. (2008). Interpreting Burrows's Delta: Geometric and probabilistic foundations. Literary and Linguistic Computing, 23(2): 131–147.

Aristotle (350 BCE). Poetics. URL *http://classics.mit.edu/Aristotle/poetics.1.1.html*.

Binongo, J. N. G. (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. Chance, 16(2): 9–17.

Börschinger, B. and Johnson, M. (2014). Exploring the role of stress in Bayesian word segmentation using Adaptor Grammars. Association for Computational Linguistics.

Bosch, R. A. and Smith, J. A. (1998). Separating hyperplanes and the authorship of the disputed Federalist Papers. American Mathematical Monthly: 601–608.

Brennan, M. R. and Greenstadt, R. (2009). Practical Attacks Against Authorship Recognition Techniques. In IAAI.

Burrows, J. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing, 17(3): 267–287.

Burrows, J. (2003). Questions of Authorship: Attribution and Beyond A Lecture Delivered on the Occasion of the Roberto Busa Award ACH-ALLC 2001, New York. Computers and the Humanities, 37(1): 5–32.

Burrows, J. (2005). Who wrote Shamela? Verifying the authorship of a parodic text. Literary and linguistic computing, 20(4): 437–450.

Burrows, J. (2007). All the way through: Testing for authorship in different frequency strata. Literary and Linguistic Computing, 22(1): 27–47.

Burrows, J. F. (1987). Word-patterns and story-shapes: The statistical analysis of narrative style. Literary and linguistic Computing, 2(2): 61–70.

Collins, J., Kaufer, D., Vlachos, P., Butler, B., and Ishizaki, S. (2004). Detecting collaborations in text comparing the authors' rhetorical language choices in the Federalist Papers. Computers and the Humanities, 38(1): 15–36.

Dillon, B., Dunbar, E., and Idsardi, W. (2013). A single-stage approach to learning phonological categories: Insights from Inuktitut. Cognitive Science, 37: 344–377.

Doyle, G. and Levy, R. (2013). Combining multiple information types in Bayesian word segmentation. In HLT-NAACL. Citeseer, pp. 117–126.

Eder, M. (2010). Does size matter? Authorship attribution, small samples, big problem. Proceedings of Digital Humanities: 132–135.

Eder, M. (2013). Mind your corpus: Systematic errors in authorship attribution. Literary and Linguistic Computing: fqt039.

Eder, M. and Rybicki, J. (2011). Stylometry with R. In Digital Humanities 2011: Conference Abstracts. Citeseer, pp. 308–311.

Feldman, N., Griffiths, T., Goldwater, S., and Morgan, J. (2013). A role for the developing lexicon in phonetic category acquisition. Psychological Review, 120(4): 751–778.

Fung, G. (2003). The disputed Federalist Papers: SVM feature selection via concave minimization. In Proceedings of the 2003 Conference on Diversity in Computing. ACM, pp. 42–46.

García, A. M. and Martin, J. C. (2007). Function words in authorship attribution studies. Literary and Linguistic Computing, 22(1): 49–66.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. Literary and Linguistic Computing, 22(3): 251–270.

Holmes, D. I. and Forsyth, R. S. (1995). The Federalist revisited: New directions in authorship attribution. Literary and Linguistic Computing, 10(2): 111–127.

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

Hoover, D. L. (2004). Testing Burrows's delta. Literary and Linguistic Computing, 19(4): 453–475.

Iqbal, F., Binsalleeh, H., Fung, B. C., and Debbabi, M. (2010). Mining writeprints from anonymous e-mails for forensic investigation. Digital Investigation, 7(1): 56–64.

Iqbal, F., Hadjidj, R., Fung, B. C., and Debbabi, M. (2008). A novel approach of mining write-prints for authorship attribution in e-mail forensics. Digital Investigation, 5: S42–S51.

Jockers, M. L. (2013). Testing authorship in the personal writings of Joseph Smith using NSC classification. Literary and Linguistic Computing, 28(3): 371–381.

Juola, P. and Baayen, R. H. (2005). A controlled-corpus experiment in authorship identification by cross-entropy. Literary and Linguistic Computing, 20(Suppl): 59–67.

Kestemont, M., Moens, S., and Deploige, J. (2015). Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux. Digital Scholarship in the Humanities, 30(2): 199–224.

Krishnapuram, B., Figueiredo, M., Carin, L., and Hartemink, A. (2005). Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27: 957–968.

Li, J., Zheng, R., and Chen, H. (2006). From fingerprint to writeprint. Communications of the ACM, 49(4): 76–82.

Lučić, A. and Blake, C. L. (2015). A syntactic characterization of authorship style surrounding proper names. Digital Scholarship in the Humanities, 30(1): 53–70.

Luyckx, K. and Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. Literary and Linguistic Computing, 26(1): 35–55.

Milic, L. T. (1966). The next step. Computers and the Humanities, 1(1): 3–6.

Mosteller, F. and Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. Journal of the American Statistical Association, 58(302): 275–309.

Mosteller, F. and Wallace, D. L. (1964). Applied Bayesian and classical inference: The case of the Federalist Papers. Springer Science & Business Media.

Oakes, M. P. (2004). Ant colony optimisation for stylometry: The Federalist Papers. In Proceedings of the 5th International Conference on Recent Advances in Soft Computing. pp. 86–91.

Pearl, L. and Steyvers, M. (2012). Detecting Authorship Deception: A Supervised Machine Learning Approach Using Author Writeprints. Literary and Linguistic Computing, 27(2): 183–196.

Price, L. (2000). The Anthology and the Rise of the Novel: From Richardson to George Eliot. Cambridge University Press.

Rokeach, M., Homant, R., and Penner, L. (1970). A value analysis of the disputed Federalist Papers. Journal of Personality and Social Psychology, 16(2): 245.

Rudman, J. (2005). The non-traditional case for the authorship of the twelve disputed Federalist Papers: A monument built on sand. Proceedings of ACH/ALLC 2005.

Savoy, J. (2013). Authorship attribution based on a probabilistic topic model. Information Processing & Management, 49(1): 341–354.

Savoy, J. (2015). Estimating the probability of an authorship attribution. Journal of the Association for Information Science and Technology.

Smith, J. B. (1989). Computer criticism. Literary Computing and Literary Criticism: Theoretical and Practical Essays on Theme and Rhetoric: 13–44.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology, 60(3): 538–556.

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

Tweedie, F. J., Singh, S., and Holmes, D. I. (1996). Neural network applications in stylometry: The Federalist Papers. Computers and the Humanities, 30(1): 1–10.

van Dalen-Oskam, K. (2014). Epistolary voices. The case of Elisabeth Wolff and Agatha Deken. Literary and Linguistic Computing: fqu023.

Vermeule, B. (2011). Why do we care about literary characters? JHU Press.

Zhao, Y. and Zobel, J. (2005). Effective and scalable authorship attribution using function words. In Information Retrieval Technology. Springer, pp. 174–189.

Zhao, Y., Zobel, J., and Vines, P. (2006). Using relative entropy for authorship attribution. In Information Retrieval Technology. Springer, pp. 92–105.

Zunshine, L. (2010). Introduction to Cognitive Cultural Studies. JHU Press.

A   Potential Features in Character Writeprints

Table 7: Potential character-level features available to the SMLR classifier for character writeprints. The feature name, description, number of individual features of this kind, and implementation are provided. One or more of the following is provided in parentheses for each feature: an example of that feature or a description of that feature.

Feature	Description	#	Implementation
alphabetic characters	all letters		
digits	all digits 0-9	3	$\frac{\#\ char\ tokens}{total\ \#\ char\ tokens}$
punctuation	all punctuation marks		
word length	average length of words	1	$\frac{\#\ char\ tokens}{\#\ word\ tokens}$
punctuation	apostrophes, colons, commas, double quotation marks, ellipses, em dashes, en dashes, exclamation marks, forward slashes, interrobangs, multiple punctuation (*!!*), parentheses, periods, questions marks, semicolons, single quotation marks, square brackets	17	$\frac{\#\ punct\ tokens}{total\ \#\ punct\ tokens}$
total characters	total # of characters	1	# character tokens

Table 8: Potential word-level features available to the SMLR classifier for character writeprints. The feature name, description, number of individual features of this kind, and implementation are provided. One or more of the following is provided in parentheses for each feature: an example of that feature and/or part-of-speech tags used to identify that feature.

Feature	Description	#	Implementation
contractions	*won't, can't*, etc.	1	$\frac{\#\ contracted\ words}{\#\ word\ tokens}$
foreign words	foreign words (*FW*)	1	$\frac{\#foreign\ words}{\#word\ tokens}$
hyphenated words	*ever-to-be-revered*, etc.	1	$\frac{\#\ hyphenated\ word\ tokens}{\#\ word\ tokens}$
lexical diversity[6]	$\frac{word\ types}{word\ tokens}$	1	$\frac{\#\ word\ types}{\#\ word\ tokens}$
total words	# total words	1	total # word tokens

Table 9: Potential content syntactic category features available to the SMLR classifier for character writeprints. The feature name, description, number of individual features of this kind, and implementation are provided. One or more of the following is provided in parentheses for each feature: an example of that feature and/or part-of-speech tags used to identify that feature.

Feature	Description	#	Implementation
adjectives	all adjectives (*good*), comparative adjectives (*JJR*), superlative adjectives (*JJS*)	3	$\frac{\# \ adj \ tokens}{\# \ word \ tokens}$
adverbs	all adverbs (*really*), basic adverbs (*RB*), comparative adverbs (*RBR*), superlative adverbs (*RBS*)	4	$\frac{\# \ adv \ tokens}{\# \ word \ tokens}$
cardinal numbers	*one, two, three*, etc.	1	$\frac{\# \ cardinal \ tokens}{\# \ word \ tokens}$
interjections	all interjections (*UH*)	1	$\frac{\# \ interjection \ tokens}{\# \ word \ tokens}$
nouns	all nouns, plural nouns (*NNS*), plural proper nouns (*NNPS*), singular or mass nouns (*NN*), singular proper nouns (*NNP*)	5	$\frac{\# \ noun \ tokens}{\# \ word \ tokens}$
ordinal numbers	*first, second, third*, etc.	1	$\frac{\# \ ordinal \ tokens}{\# \ word \ tokens}$
possessives	Clarissa*'s* (*POS*)	1	$\frac{\# \ poss \ tokens}{\# \ word \ tokens}$
pronouns	1st person pronouns (*I*), 2nd person pronouns (*you*), 3rd person pronouns (*he*), demonstrative pronouns (*this*), personal pronouns (*PRP*), possessive pronouns (*PRP\$*), relative pronoun (*which*)	7	$\frac{\# \ pronoun \ tokens}{\# \ word \ tokens}$
verbs	all verbs, gerund or present participle (*VBG*), non-finite verbs (*VB*), past participle (*VBN*), past tense verbs (*VBD*)	5	$\frac{\# \ verb \ tokens}{\# \ word \ tokens}$

Table 10: Potential functional syntactic category features available to the SMLR classifier for character writeprints. The feature name, description, number of individual features of this kind, and implementation are provided. One or more of the following is provided in parentheses for each feature: an example of that feature, part-of-speech tags used to identify that feature, or a collection of tokens comprising that feature.

Feature	Description	#	Implementation
coordinating conjunctions	*although, and, because, but, for, nor, or, since, so, though, unless, while, yet*	1	$\frac{\#\ coordinating\ conj}{\#\ word\ tokens}$
determiners	additive (*more*), alternative (*another, other, somebody else, different*), articles (*a, an, the*), disjunctive (*either, neither*), distributive (*each, every*), elective (*any, either, whichever*), equative (*same*), evaluative (*such, so, that*), exclamative (*what* cheek), existential (*some, any*), interrogative & relative (*which, what, whichever, whatever*), maximal & minimal (*most, least*), negative (*no, neither*), paucal (*a few, a little, some*), personal (*we* friends, *you* scoundrels), quantifiers (*all, few, many, several, some, each, every, each, any, no, a lot of, much*), subtractive (*less, fewer*), sufficiency (*enough, sufficient, plenty*), universal (*all, both, every*)	19	$\frac{\#\ det\ tokens}{\#\ word\ tokens}$
function words	all function words: articles (*a, an, the*), copula be, determiners (*DT*), expletives (*EX*), infinitival to, prepositions, personal pronouns, possessive pronouns (*PRP*), possessives (*PRP$*), relative pronouns (*which*)	1	$\frac{\#\ function\ tokens}{\#\ word\ tokens}$
infinitival to	*to* go	1	$\frac{\#\ infin\ to\ tokens}{\#\ word\ tokens}$
prepositions	*with*, etc.	1	$\frac{\#\ prep\ tokens}{\#\ word\ tokens}$
sentence connectors	occurring at the beginning or end of sentences: *also, anyway, as, besides, finally, first, furthermore, hence, however, in addition, last but not least, lastly, moreover, nevertheless, on the other hand, otherwise, second, so, still, then, thus, too, well, yet*	1	$\frac{\#\ sent\ connectors}{\#\ word\ tokens}$

Table 11: Potential syntactic structure features available to the SMLR classifier for character writeprints. The feature name, description, number of individual features of this kind, and implementation are provided. One or more of the following is provided in parentheses for each feature: an example of that feature, part-of-speech tags used to identify that feature, or phrase-structure tags used to identify that feature.

Feature	Description	#	Implementation
avg phrase length	adjective phrases (*ADJP*), adverb phrases (*ADVP*), conjunction phrases (*CONJP*), noun phrases (*NP*), parenthetical phrases (*(...)*), preposition phrases (*PP*), quantifier phrases (QP), verb phrases (*VP*)	8	$\frac{\#\ word\ tokens\ in\ phrase\ type}{total\ \#\ phrase\ type}$
avg sentence length	sentences, exclamations, questions	3	$\frac{\#\ word\ tokens\ in\ sent\ type}{total\ \#\ sent\ type}$
clauses	*She laughed*, and then *she cried*	1	$\frac{\#\ clause\ tokens}{\#\ sentences}$
embedded clauses	the ones *that I like*	1	$\frac{\#\ emb\ cl\ tokens}{\#\ clauses}$
exclamations	*What a surprise!*	1	$\frac{\#\ excl\ tokens}{total\ \#\ sentences}$
fragments	fragments (*FRAG*)	1	$\frac{\#\ fragment\ tokens}{\#\ sentences}$
imperatives	*Do this now!*	1	$\frac{\#\ imperative\ tokens}{\#\ sentences}$
main clauses	*I think* she's right.	1	$\frac{\#\ main\ cl\ tokens}{\#\ clauses}$
passives	*was kissed*, etc.	1	$\frac{\#\ passive\ tokens}{\#\ sentences}$
phrases (rel freq)	adjective phrases (*ADJP*), adverb phrases (*ADVP*), conjunction phrases (*CONJP*), noun phrases (*NP*), preposition phrases (*PP*), quantifier phrases (*QP*), verb phrases (*VP*)	7	$\frac{\#\ phrase\ type\ tokens}{\#\ phrases}$
phrases	all phrases (*ADJP, ADVP, CONJP, NP, PP, QP, VP, WHADJP, WHADVP, WHNP, WHPP*)	1	# phrase tokens
questions	*Was that a surprise?*	1	$\frac{\#\ questions\ tokens}{\#\ sentences}$
sentences	total sentences	1	# sentence tokens
wh-phrases	wh-adjective phrases (*how hot, WHADJP*), wh-adverb phrases (*how, WHADVP*), wh-noun phrases (*what, WHNP*), wh-preposition phrases (*to what, WHPP*)	4	$\frac{\#\ wh-phrase\ tokens}{\#\ phrases}$

Table 12: Potential formatting features available to the SMLR classifier for character writeprints. The feature name, description, number of individual features of this kind, and implementation are provided. An example or description of each feature is provided.

Feature	Description	#	Implementation
all capitals	*LIKE*, etc.	1	$\frac{\text{\# caps word tokens}}{\text{\# word tokens}}$
italicized phrases	average length of italicized phrases	1	$\frac{\text{\# ital word tokens}}{\text{\# italicized phrases}}$
parenthetical words	This seems (*decidedly*) interesting, etc.	1	$\frac{\text{\# paren word tokens}}{\text{\# word tokens}}$
italicized words	*like*, etc.	1	$\frac{\text{\# ital word tokens}}{\text{\# word tokens}}$

Table 13: Potential domain-specific semantic features available to the SMLR classifier for character writeprints. The feature name, description, number of individual features of this kind, and implementation are provided. One or more of the following is provided in parentheses for each feature: an example of that feature or a collection of tokens comprising that feature.

Feature	Description	#	Implementation
endearments	*affectionate(ly), beloved, (my) bird, charmer, darling, (my) dear, dearest, dearly, (ever-)affectionate, (my) good, goodness, honey, lamb, (my) love, (my) pet, sincere, sincerity, (my) sweet*	1	$\frac{\text{\# endearment tokens}}{\text{\# word tokens}}$
epistolary	*correspondence, letter(s), paper(s), parcel(s), pen(s), receipt(s), resume, write, written, writing, wrote*	1	$\frac{\text{\# epistolary tokens}}{\text{\# word tokens}}$
familial	*aunt, brother, child, daughter, family, father, grandfather, mamma, maternal, mother, papa, paternal, sister, son, uncle*	1	$\frac{\text{\# familial tokens}}{\text{\# word tokens}}$
propriety	*compliment, congratulate, excuse me, grateful, manners, obliged, politeness, propriety, thank you*	1	$\frac{\text{\# propriety tokens}}{\text{\# word tokens}}$
"remarkable" adverbs	*best, better, quite, so, soon, sooner, soonest, too, well*	1	$\frac{\text{\# remarkable tokens}}{\text{\# word tokens}}$
titles	*Dr, Esq, Jr, Lord, Madam, Miss, Ms, Mr, Mrs, Reverend, Sir, Sr*	1	$\frac{\text{\# title tokens}}{\text{\# word tokens}}$
writing	*authorship, compose, composition, correspondence, drop a line, indite, ink, letter(s), paper(s), parcel, pen(s), penning, receipt, resume, spell, words, write, (piece of) writing, written material*	1	$\frac{\text{\# writing tokens}}{\text{\# word tokens}}$

B Detailed Character Precision and Recall Scores

Below we show the details of the confusion matrices as well as the precision and recall scores used to calculate the F-scores reported in the main text for each character. Precision and recall scores range between 0.0 and 1.0.

Table 14: Specific confusion matrix values and accompanying precision and recall scores for Clarissa Harlowe's letters.

		Labeled		Recall
		Clarissa	non-Clarissa	
True	Clarissa	136	108	0.557
	non-Clarissa	128	416	
	Precision	0.515		

Table 15: Specific confusion matrix values and accompanying precision and recall scores for Lovelace's letters.

		Labeled		Recall
		Lovelace	non-Lovelace	
True	Lovelace	109	89	0.551
	non-Lovelace	84	506	
	Precision	0.565		

Table 16: Specific confusion matrix values and accompanying precision and recall scores for John Belford's letters.

		Labeled		Recall
		Belford	non-Belford	
True	Belford	42	62	0.404
	non-Belford	66	618	
	Precision	0.389		

Table 17: Specific confusion matrix values and accompanying precision and recall scores for Anna Howe's letters.

		Labeled		Recall
		Anna	non-Anna	
True	Anna	18	60	0.231
	non-Anna	70	640	
	Precision	0.205		