Finding Characteristic Features in Stylometric Analysis

Klaussner, Carmen; Nerbonne, John; Çöltekin, Çağri

Published in:
Digital Scholarship in the Humanities

DOI:
10.1093/llc/fqv048

Finding Characteristic Features in Stylometric Analysis

Carmen Klaussner
Trinity College Dublin, Ireland

John Nerbonne
University of Groningen, The Netherlands and University of Freiburg, Germany

Çağrı Çöltekin
University of Groningen, The Netherlands

Abstract

The usual focus in authorship studies is on authorship attribution, i.e. determining which author (of a given set) wrote a piece of unknown provenance. The usual setting involves a small number of candidate authors, which means that the focus quickly revolves around a search for features that discriminate among the candidates. Whether the features that serve to discriminate among the authors are characteristic is then not of primary importance. We respectfully suggest an alternative in this article, namely a focus on seeking features that are characteristic for an author with respect to others. To determine an author's characteristic features, we first seek elements that he or she uses consistently, which we therefore regard as 'representative', but we likewise seek elements which the author uses 'distinctively' in comparison to an opposing author. We test the idea on a task recently proposed that compares Charles Dickens to both Wilkie Collins and a larger reference set comprising several authors' works from the 18th and 19th century. We then compare the use of representative and distinctive features to Burrows' 'Delta' and Hoovers' 'CoV Tuning'; we find that our method bears little similarity with either method in terms of characteristic feature selection. We show that our method achieves reliable and consistent results in the two-author comparison and fair results in the multi-author one, measured by separation ability in clustering.

Correspondence:
John Nerbonne, CLCG, P.O.Box 716, University of Groningen, 9700 AS Groningen, The Netherlands.
E-mail: j.nerbonne@rug.nl

1 Introduction

This article suggests a novel, complementary focus in stylometry, i.e. trying to identify characteristic features of authors rather than focusing on discriminating among authors, which is the common task in authorship attribution. The latter has served to focus scholars on a task with clear success criteria, certainly an achievement, but we suspect that its focus on finding discriminating features leads to an overemphasis on unusual features rather than characterizations of what is general and consistent about an author's style. We thus ask with others 'If you can tell authors apart, have you learned anything about them?' (Craig, 1999). Concretely we try to identify words that Dickens uses with a consistent frequency

Digital Scholarship in the Humanities, Vol. 30, Supplement 1, 2015.

doi:10.1093/llc/fqv048 Advance Access published on 9 October 2015

throughout a selection of his writings and which are used differently by other authors. We think that the approach might be used to analyze syntactic features, too, but we will not try to show that.

The field of stylometry in authorship studies has undergone considerable change in the course of the 20th century, whose beginning marked the tentative introduction of new measures to the field, heralding the rise of non-traditional, quantitative techniques to be established alongside the then predominant traditional methods (e.g. manuscript provenance or dating of materials). In the interest of space, we shall not summarize that history here, referring instead to excellent recent surveys (Stamatatos, 2009; Oakes, 2014).

Since Burrows' work is a touchstone for many, we discuss it here specifically and compare our proposal to his work in more detail below. Burrows' 'Delta' (Burrows, 2002) was designed for authorship attribution, seeking the most likely authorial candidate for a given document from a set of authors based on differences between z-scores of high-frequency items. Delta is usually applied to the 800–1,000 most frequent words, i.e. the highest-frequency stratum. This is an advantage since high-frequency words are likely to be encountered in most documents. But note that highly variable features could be useful for the task of identifying an author if they happened to occur almost exclusively in just one author's works, but we would not regard them as characteristic since they are not used consistently. Burrows' 'Iota' and 'Zeta' (Burrows, 2005, 2007; Hoover, 2007) investigate words in middle-range and low-range frequency strata, and they look for words appearing consistently in one author's works and less frequently to not at all (Iota) in the works of others. More recently, Hoover introduced 'CoV Tuning', that uses the Coefficient of Variance to detect those frequent features that are most variable over a multi-author corpus (Hoover, 2014).[1]

We introduce a new technique, Representativeness and Distinctiveness (RD), focusing on finding style markers that are used consistently in the works of one author and differently from that of others. Concretely, we try to detect Charles Dickens' style presented by Tabata (2012), who used Random Forest classification. We compare our results to Tabata's in Section 4.3.

The remainder of this article is organized as follows; we begin by introducing and further motivating RD in Section 2 in the context of style analysis. Section 3 gives an overview of the data; Section 4 continues by first exemplifying our technique's application to an actual task and subsequently comparing it to other methods in the field. We close the discussion in Section 5.

2 Finding Characteristic Features

Rather than focusing exclusively on identifying stylistic features that discriminate among authors, we first seek features that an author uses consistently in his work, calling these features representative, and turn to distinctive features in a second step. In dialectology, where these methods were first used, we note, e.g. that the word used for the storage space in a car is fairly consistently call a 'boot' throughout the UK and similarly that the words 'cot' and 'caught' rhyme on the Eastern seaboard of the USA. This makes them representative. We do not have atomistic data of this detail in stylometry, where there is a long and serious tradition of looking first to word frequencies as style markers. We therefore focus on word frequencies here, but we might also have examined the frequencies of word bigrams or sequences of part-of-speech tags.

In order to identify what is consistent in an author's style, we consider not only the very highest strata of frequent words (i.e. 1–800), but rather a larger set (i.e. 1–5,000). The aim of this is to find features with a very even distribution over an author's works; those used very frequently and those used less frequently. Naturally, very infrequent features will suffer the instability problems associated with sparse data, so we do not imagine using them effectively.

Distinctive features are always identified with respect to a set of comparable authors, and they are simply the features used differently by the candidate under examination and the comparable set.

We turn now to a more formal introduction of RD and further explanation of how it can be used in stylometry. More specific applications of the

method are presented in Section 4, where we test the method in two different settings.

2.1 Representativeness and Distinctiveness

Representativeness and Distinctiveness were introduced in dialectology (Wieling and Nerbonne, 2011), with the goal of detecting linguistic features that mark the speakers of a particular dialect in contrast to others. In the original paper, it is used to detect characteristic features (e.g. lexical items), that differ little within the target group of geographical sites (and may therefore be regarded as 'representative') and differ considerably more outside that group (so that they are also 'distinctive' with respect to the other group). It was later extended to function with numerical measures (Prokić *et al.*, 2012), and since we will analyze frequency, we will focus on that extension.

In authorship analysis, we examine the words extracted from an author's documents compared to documents by another group of authors ($\sim$ the reference set). More exactly, we examine the frequency distribution of the author's vocabulary, as it is used across the range of documents (or text segments). The technique begins by identifying which feature frequencies are consistent over the target author's document set. Afterwards, it selects those consistent and thus 'representative' features of that author that are also 'distinctive' with respect to those documents in the (contrasting) reference set.

We assume a set of documents from an author under investigation, D_{in} as well as a set of contrasting documents, D_{ex}, which we need if we are to identify distinctive features. We may also refer to D, $D = D_{\text{in}} \cup D_{\text{ex}}$, the union of the two sets. We assume moreover a distance function 'diff', which for a given feature f, returns the distance between a pair of documents with respect to f.

The formal definition of the Representativeness of a particular feature f for a document set D_{in} (belonging to the target author) is then based on the mean distance of the documents in D_{in} with respect to f:

$$\overline{d_f^{D_{\text{in}}}} = \frac{2}{|D_{\text{in}}|^2 - |D_{\text{in}}|} \sum_{d,d' \in D_{\text{in}}, d \neq d'} \text{diff}_f(d, d') \quad (1)$$

where the fraction before the summation is based on the number of non-identical pairs in the set D_{in}.

Naturally, we also need to know the average distance between pairs of documents, where the first comes from D_{in} and the second from D_{ex}. These allow us to compare the target author to others:

$$\overline{d_f^{D}} = \frac{1}{|D_{\text{in}} \times D_{\text{ex}}|} \sum_{d \in D_{\text{in}}, d' \in D_{\text{ex}}} \text{diff}_f(d, d') \quad (2)$$

where we assume, as noted above, that $D = D_{\text{in}} \cup D_{\text{ex}}$. We implicitly appeal to the assumed definition in order to suppress the reference to two document sets on the left-hand side of the definition. We deliberately collect feature frequencies not only when they are greater than those in the reference set, but also when they are less.

In order to determine features both representative of a particular author as well as distinctive with respect to other authors, we normalize the average values defined in eq. 1 and eq. 2 above.

$$\text{Repr}_f(D_{\text{in}}) = -\frac{\overline{d_f^{D_{\text{in}}}} - \overline{d_f}}{sd(d_f)} \quad (3)$$

$$\text{Dist}_f(D) = \frac{\overline{d_f^{D}} - \overline{d_f}}{sd(d_f)} \quad (4)$$

where $\overline{d_f}$ is the mean difference between all documents within the document set D, $D = D_{\text{in}} \cup D_{\text{ex}}$, with respect to the feature f, where $sd(d_f)$ is the standard deviation of differences between all documents in the document set with respect to f, and where we again implicitly assume that $D = D_{\text{in}} \cup D_{\text{ex}}$. Note that Repr is defined as the negative of the normalized $\overline{d_f^{D_{\text{in}}}}$, since smaller internal differences mean more consistent features. The normalization step also makes sure that Representativeness not only measures consistent features within an author's documents, but that it also compares them to the rest of the documents. Hence, only the features that are exceptionally consistent within the target author's documents in comparison to the other documents will receive higher Repr scores. Similarly, the Dist measure does not just select highly variable features in the language, but will score highly those features whose

use contrasts between the target author's documents and the reference set.

We define the features that are both representative and distinctive as the 'characteristic features' of an author. In this article, we use the sum of Repr and Dist to obtain a single summary score representing how characteristic a feature is for the author of interest. We refer to this combined score (Repr + Dist) as the RD_f score, and refer then to $RD_f(A, B)$ or $RD_f(D_{in}, D_{ex})$. For different applications, other combinations of Repr and Dist may be more appropriate.

2.2 Distinctiveness in comparing only two authors

The RD measure, as defined above, compares texts written by an author with a reference set typically comprising many other authors. In some of the experiments (reported in Section 4.1), we present results comparing only two authors. This subsection discusses the interpretation of the measures in the two-author setting and clarifies further properties of the RD_f score.

In the two-author setting, we have two sets of documents, one belonging to author A and the other to author B (or to D_{in}, D_{ex}), respectively. We first consider the case where the same feature is representative in both authors' works. If the feature is used consistently at the same rate by both authors, it will be representative for both individually, but not distinctive. If it is used consistently by both but at different rates, then it may score well in Distinctiveness depending on the size of the difference. So, representative features need not result in high RD_f scores.

The RD measure may be symmetric, e.g. when feature f is representative in set D_{in} because it occurs with a consistently high frequency. If the same feature f is also representative in the opposing set, D_{ex}, but with a low frequency, then f will be representative and distinctive for both sets, and $RD_f(A, B) = RD_f(B, A)$.

But the measure may be asymmetric, so that $RD_f(A, B) \neq RD_f(B, A)$, if e.g. the feature is highly representative in A but not B. This means that a representative and distinctive feature for the candidate set D_{in}, may be unrepresentative for set D_{ex} because its frequencies may vary too much in the documents in D_{ex}. Although this feature is not representative for D_{ex}, it may still be distinctive in D_{in} with respect to D_{ex}, because it is used with consistent frequency in D_{in} but not in D_{ex}.

Thus, high RD_f scores indicate consistent frequencies within the target author's documents that may either be inconsistent or be consistently different in the reference set. The values obtained do not reveal whether an author consistently avoided or preferred a particular feature. A given feature f may be scored highly relevant for both authors, so that $RD_f(A, B) \approx RD_f(B, A)$, meaning one uses it consistently less often than the other, rendering it a good separator for the two authors.

2.2.1 *General properties*

From a performance point of view, the more features (or documents) one considers, the more expensive the computations will be, since the methods require pairwise comparisons of all documents for each individual feature.[2]

3 Data

In this section, we introduce the data sets used in all the experiments reported on in Section 4. The exact composition of the data sets was motivated by a study by Tabata (2012), where Charles Dickens was contrasted both with contemporary writer Wilkie Collins in a two-author comparison and with a larger reference set comprising different authors from the 18th and 19th century and thus a reference for the 'average' writing style of that time. For all experiments, we consider the data sets proposed by Tabata (2012), namely a set consisting of twenty-four texts by Dickens and Collins each (shown in Tables 1 and 2, respectively).[3] Thus, while the data set for the first experiment here is the same as used by Tabata (2012), we assembled the data for the second experiment ourselves; these contain the same texts for Dickens as in the first experiment, while the reference set in this second case contains fifty-five texts by sixteen different authors. The texts are shown in Tables 3 and 4. This data set was preprocessed by removing all punctuation, but retaining contractions and compounds

Table 1 Dickens' texts

Author	Texts	Year
Dickens	Sketches by Boz	1833–36
Dickens	The Pickwick Papers	1836–37
Dickens	Other Early papers	1837–40
Dickens	Oliver Twist	1837–39
Dickens	Nicholas Nickleby	1838–39
Dickens	Master Humphrey's Clock	1840–41
Dickens	The Old Curiosity Shop	1840–41
Dickens	Barnaby Rudge	1841
Dickens	American Notes	1842
Dickens	Martin Chuzzlewit	1843–44
Dickens	Christmas books	1843–48
Dickens	Pictures From Italy	1846
Dickens	Dombey and Son	1846–48
Dickens	David Copperfield	1849–50
Dickens	A Child's History of England	1851–53
Dickens	Bleak House	1852–53
Dickens	Hard Times	1854
Dickens	Little Dorrit	1855–57
Dickens	Reprinted Pieces	1850–56
Dickens	A Tale of Two Cities	1859
Dickens	The Uncommercial Traveller	1860–69
Dickens	Great Expectations	1860–61
Dickens	Our Mutual Friend	1864–65
Dickens	The Mystery of Edwin Drood	1870

Table 2 Collins' texts

Author	Texts	Year
Collins	Antonina	1850
Collins	Rambles Beyond Railways	1851
Collins	Basil	1852
Collins	Hide and Seek	1854
Collins	After Dark	1856
Collins	A Rogue's Life	1856–57
Collins	The Queen of Hearts	1869
Collins	The Woman in White	1860
Collins	No Name	1862
Collins	Armadale	1866
Collins	The Moonstone	1868
Collins	Man and Wife	1870
Collins	Poor Miss Finch	1872
Collins	The New Magdalen	1873
Collins	The Law and the Lady	1875
Collins	The Two Destinies	1876
Collins	The Haunted Hotel	1878
Collins	The Fallen Leaves	1879
Collins	Jezebel's Daughter	1880
Collins	The Black Robe	1881
Collins	I Say No	1884
Collins	The Evil Genius	1886
Collins	Little Novels	1887
Collins	The Legacy of Cain	1888

Table 3 Eighteenth-century texts

Author	Texts	Year
Defoe	Captain Singleton	1720
Defoe	Journal of Prague Year	1722
Defoe	Military Memoirs of Capt. George Carleton	1728
Defoe	Moll Flanders	1724
Defoe	Robinson Crusoe	1719
Fielding	A journey from this world to the next	1749
Fielding	Amelia	1751
Fielding	Jonathan Wild	1743
Fielding	Joseph Andrews I&II	1742
Fielding	Tom Jones	1749
Goldsmith	The Vicar of Wakefield	1766
Richardson	Clarrissa I - IX	1748
Richardson	Pamela	1740
Smollett	Peregrine Pickle	1752
Smollett	Travels through France and Italy	1766
Smollett	The Adventures of Ferdinand Count Fathom	1753
Smollett	Humphrey Clinker	1771
Smollett	The Adventures of Sir Launcelot Greaves	1760
Smollett	The Adventures of Roderick Random	1748
Sterne	A Sentimental Journey	1768
Sterne	The Life and Opinions of Tristram Shandy	1759–67
Swift	A Tale of a Tub	1704
Swift	Gulliver's Travels	1726
Swift	The Journal to Stella	1710–13

and transforming the data by computing relative frequencies multiplied by 100. Finally, we remove document-specific features over the whole corpus by probing whether a term appears in at least two-third of the documents and discarding it otherwise.

We note that both data preparation steps—limiting features to the most frequent ones and filtering those that do not appear regularly—serve to increase the chance of using features we would call 'representative'. Eliminating infrequent features reduces noise and increases the chance of settling on statistically stable elements.

4 Experiments

In this section, we begin by considering the task proposed by Tabata (2012), i.e. that of determining Dickens' characteristic features. We do this by first comparing his works to his contemporary Collins and then to a reference corpus; this is done in

Table 4 Nineteenth-century texts

Author	Texts	Year
Brontë, A.	Agnes Grey	1847
Austen	Emma	1815
Austen	Mansfield Park	1814
Austen	Pride and Prejudice	1813
Austen	Northanger Abbey	1803
Austen	Sense and Sensibility	1811
Austen	Persuasion	1816–18
Brontë, C.	The Professor	1857
Brontë, C.	Villette	1853
Brontë, C.	Jane Eyre	1847
Brontë, E.	Wuthering Heights	1847
Eliot	Daniel Deronda	1876
Eliot	Silas Marner	1861
Eliot	Middlemarch	1871–72
Eliot	The Mill on the Floss	1860
Eliot	Brother Jacob	1864
Eliot	Adam Bede	1859
Gaskell	Cranford	1851–53
Gaskell	Sylvia's Lovers	1863
Gaskell	Mary Barton	1848
Thackeray	Vanity Fair	1848
Thackeray	Barry Lyndon	1844
Trollope	Doctor Thorne	1857
Trollope	Barchester Towers	1857
Trollope	The Warden	1855
Trollope	Phineas Finn	1869
Trollope	Can You Forgive Her	1865
Trollope	The Eustace Diamonds	1873
Collins	After Dark	1882
Collins	The Moonstone	1868
Collins	The Woman in White	1859

Section 4.1 and Section 4.2, respectively. In order to analyze the extent to which the method proposed here is different from the machine-learning technique used by Tabata (2012), we compare our results to Tabata's in Section 4.3. Further, we consider comparisons both to Burrows' well-established method (Burrows' Delta in Section 4.4), as well as to a more recently introduced technique (Hoover's CoV Tuning in Section 4.5).

4.1 Dickens versus Collins

Charles Dickens is perceived to have a somewhat unique style that sets his pieces apart from his contemporaries (Mahlberg, 2007). This makes him a good subject for style analysis, as there are likely to be features that distinguish him from others. Thus, Dickens has been focus of numerous stylistic analyses (Mahlberg, 2007; Craig and Drew, 2011; Tabata, 2012). The study presented by Mahlberg (2007) describes a work aimed at introducing corpus linguistics methods to extract key word clusters (sequences of words), that can then be interpreted more abstractly in a second step. The study focuses on twenty-three texts by Dickens in comparison to a 19th-century reference corpus, containing twenty-nine texts by various authors and thus a sample of contemporary writing. According to Mahlberg, Dickens shows a particular affinity for using 'Body Part' clusters: e.g. 'his hands in his pockets', which is interpreted as an example of Dickens' individualization of his characters. Although this use is not unusual for the time, the rate of use in Dickens is remarkable, as Dickens, for instance, links a particular bodily action to a character more than average for the 19th century. The phrase 'his hands in his pockets', for instance, occurs ninety times and in twenty texts of Dickens, compared to thirteen times and eight texts in the 19th-century reference corpus.

Mahlberg concludes that the identification of body part clusters provides further evidence of the importance of body language in Dickens. Thus, frequent clusters can be an indication of what function (content) words are likely to be or not be among Dickens' discriminators, in this case, we would expect there to be examples of body parts, such as 'face', 'eyes', and 'hands'.

For the comparison between Dickens and Collins, we consider the same data used by Tabata (2012). The combined data set contains twenty-four documents each for the two authors, for which the first ~5,000 most frequent words were extracted. For evaluation, we return to the authorship evaluation task, since, after all, characteristic words should serve to discriminate between authors, but we take care to attend to the words responsible for the discrimination as well.

We use five-fold cross-validation and subsequent clustering of documents which we evaluate using the 'Adjusted Rand Index (ARI)' (Hubert and Arabie, 1985), where 0 is the expected (chance) value and 1 perfect overlap with a (gold) standard. The input features for clustering are selected by considering the shared items of the *n*-highest rated features of

Table 5 Results for five-fold cross-validation for discriminating in the Dickens/Collins set, with 'Input' referring to the number of features selected from the (top of the) lists of the two authors' representative and distinctive features and 'Shared' to the number of those input features shared by both. The shared features are used in clustering. Results for clustering on the entire set/test set are shown in the other columns.

Feature number		ARI									
		Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
Input	Shared	Full	Test	Full	Test	Full	Test	Full	Test	Full	Test
100	46	0.84	0.16	1	1	0.84	1	0.84	1	1	1
200	79	0.84	0.49	0.92	1	0.84	1	0.84	1	0.84	1
300	107	0.84	0.49	0.84	1	0.84	1	0.84	1	0.84	1
400	130	0.84	0.49	0.84	1	0.84	1	0.84	1	0.84	1
500	157	0.84	0.49	0.84	1	0.84	1	0.84	1	0.84	1
1,000	305	0.84	0.49	0.84	1	0.92	1	0.84	1	0.84	1
2,000	1,045	0.84	0.49	0.84	1	0.84	1	0.84	1	0.84	1
3,000	2,188	0.84	0.49	0.92	1	0.84	0	0.84	0	0.84	1
3,250	2,509	0.84	0.16	0.92	1	0.00	0	0.84	0	0	1

the two authors, with n iterating from 100 to the total length of the feature input list in steps of fifty, e.g. 100, 150, 200, . . ., 5,000. The distance matrix was computed using the 'Manhattan' distance and subsequent clustering was performed using 'complete link' (Manning *et al.*, 2008).

Table 5 shows selected results, where 'Input' refers to the features originally selected and 'Shared' to those selected by the RD_f scores for both authors and therefore retained for clustering. For each iteration, we show the ARI for clustering on the complete data set and on the test set only. The results are very regular, even when increasing the feature input size dramatically. However, at 2,509 shared features, the accuracy decreases, and this deterioration continues in subsequent iterations. Fold 1 is considerably and consistently worse for the test set accuracy than the other folds. Upon examining its test documents, it can be observed that two unusual pieces of Collins are part of this set, *Antonina* and *Rambles Beyond Railways*, which Tabata also identified as conspicuous in Collins' works (Tabata, 2012).

Further, we can examine prominent features of the two authors in Table 6, which shows the fifteen highest-rated representative and distinctive features for each author. The six features in bold are shared by Dickens and Collins and appear among the top fifteen items based on RD_f scores. These features are thus not only distinctive, but also representative in their frequency distributions for Dickens and Collins. This means that one of them uses the item consistently more frequently than the other. Considering the consistency of results, the method is likely to be appropriate for two-author comparisons.

Table 6 Representative and distinctive scores for highest features on 300 input features in Fold 1

Dickens		Collins	
Feature	RD_f score	Feature	RD_f score
left	1.78	**upon**	1.91
letter	1.74	**though**	1.81
only	1.74	such	1.74
first	1.73	so	1.71
discovered	1.71	**only**	1.69
later	1.71	being	1.67
but	1.70	**but**	1.66
produced	1.69	much	1.65
advice	1.69	many	1.61
wait	1.68	answer	1.59
upon	1.68	very	1.59
though	1.66	and	1.57
words	1.64	**left**	1.56
future	1.64	to	1.56
news	1.63	**first**	1.53

Shared features are marked in bold.

Table 7 Results for five-fold cross-validation on the Dickens/World set, with 'Input' referring to the number of highest features selected from Dickens' and the reference corpus' representative and distinctive features and 'Shared' to the number of those input features shared by the two sets—these are used in clustering. Results for clustering on the entire set/test set are shown.

Feature number		ARI									
		Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
Input	Shared	Full	Test	Full	Test	Full	Test	Full	Test	Full	Test
100	9	–0.01	–0.08	0.76	0.51	0.03	1	0.76	–0.07	0.76	0.54
200	12	0.54	1	0.76	0.51	0.03	1	0.67	–0.07	0.76	0.54
300	27	0.03	1	0.03	0.51	0.03	1	0.67	–0.07	0.80	0.54
400	43	0.03	1	0.67	0.51	0.03	1	0.03	–0.07	0.67	0.35
500	78	0.18	1	0.03	0.32	0.03	0.22	–0.01	–0.07	0.63	0.75
1,000	407	–0.07	1	–0.03	0.51	–0.04	0.22	–0.04	0.09	–0.04	0.08

4.2 Dickens versus 'World'

In the second experiment presented by Tabata (2012), the task was to identify Dickens' style with respect to a larger reference corpus, in order to detect items that set him apart from other authors of his time rather than only Collins. Thus, we consider the same texts used in that exercise and transformed the data by computing relative frequencies and excluding words not present in at least two-third of the complete data set, which reduces it to ~4,000 input features (words).

Table 7 shows the cross-validation results for clustering Dickens vs. the reference corpus. As in the previous case, the distance matrix was computed using the 'Manhattan' distance and subsequent clustering was done using 'complete link'. In contrast to the Dickens-Collins comparison, the results are less consistent. In order to obtain a fair number of shared features, the number of input features has to be much greater than in the two-author experiment.

In the previous case, there were two pieces in the first fold's test set that are likely to have lowered the overall ARI (see above). Of course, this can happen in other trial runs based on a random five-fold cross-validation. If there are only a few documents of a given author and these are (almost) all missing from the training corpus, they are more likely to be misclassified in clustering. The test set in Fold 3 is an interesting candidate; clustering based on a higher set of features is quite low, close to the expected value of random clustering, while the test set results based on fewer features are generally quite high. The test set for this fold consists of four novels by Dickens, of all six of the novels by Austen in the data set, and one each by Smollett and Sterne and each of the Brontë sisters. Closer inspection reveals that the absolute distance between clusters is very slight for the test documents.

Clustering the complete data set shows that seven documents are misclassified—namely all three novels of Charlotte Brontë as well as one by Thackeray, Smollett, Sterne, and Dickens each. Interestingly, all of Austen's novels are correctly attributed, despite the fact that none of her works were part of the training corpus, suggesting that her style is sufficiently similar to her peers. This might also suggest that Austen is not only very consistent within her own texts, but presents a kind of 'average' of the corpus, while certain authors/works deviate more from this.

The only fold (Table 7) that behaves more regularly is Fold 5, where both the full set and the test set have mediocre to fair results, suggesting that the test documents in this case (Gaskell (1/2), Eliot (4/6), Trollope (2/6), Collins (2/3), Thackeray (1/2)) were a better reflection of the training corpus, which in fact did contain samples of these authors. Overall, one can conclude that the composition of the reference set, as well as possible prevalence of particular authors might considerably influence the selection of features.

Table 8 shows the fifteen highest-rated features for both Dickens and the reference corpus. In this

Table 8 Scores for highest features on 300 input features in Fold 5

Dickens		World	
Feature	**RD_f score**	**Feature**	**RD_f score**
corner	1.10	head	1.25
given	1.10	**corner**	1.24
quiet	1.03	old	1.19
till	0.99	legs	1.16
for	0.99	various	1.15
return	0.98	hat	1.08
pleased	0.96	shaking	0.99
however	0.96	until	0.96
entirely	0.94	looking	0.96
give	0.94	remark	0.96
use	0.93	heavily	0.92
without	0.93	returned	0.92
able	0.92	raising	0.90
cannot	0.92	behind	0.90
upon	0.92	faces	0.90

case, the scores for each are considerably lower than for Dickens and Collins in the previous experiment. This suggests that consensus over features is more difficult to attain for the larger reference set, which in turn affects the degree of Distinctiveness for Dickens (even if his features' Representativeness will be the same in this case). The number of shared items is also lower than it was previously when we considered the same number of highest features. However, among the first thirty items of both lists, there are a number of body parts, such as 'head', 'faces', and 'legs', as well as words denoting action, such as 'looking', 'shaking', and 'raising', indicating that these indeed distinguish Dickens from his contemporaries, one giving preference to these expressions, while the others are rather avoiding them. While RD cannot reveal which of these expressions Dickens himself preferred, taking into consideration previous analyses (Mahlberg, 2007; Tabata, 2012), we might tentatively conclude that he used the above more frequently than his peers.

4.3 Comparing to Tabata's Random Forests

In the following, we compare our results to the ones obtained by Tabata (2012), who used Random Forests (RF) Classification on the same two tasks we reported on in the last two sections.

4.3.1 *RF classification*

RF was first introduced by Breiman (2001) and is based on ensemble learning from a large number of decision trees randomly generated from the data set. The 'forest' is created by building each tree individually by sampling n cases (documents) at random with replacement (with n ~66% of the complete data). At each node, m predictor variables are selected at random from all the predictor variables, finally choosing the variable that provides the best split, according to some objective function (m << total number of predictor variables). A new document is classified by taking an average or weighted average or a voting majority in the case of categorical variables.

In terms of interpretability, RF classification offers more transparency than other machine-learning algorithms, in that it indicates what variables were important in classification, in the present case, which words were best in separating Dickens from Collins or from the 18th-/19th-century reference set. For both experiments in Tabata (2012), the 300 most frequent words were used as input features, yielding a list of features for Dickens and Collins each, shown in Table 9 and one for Dickens' positive and negative features when compared to the larger reference corpus, as shown in Table 10.

4.3.2 *Characteristic feature comparison*

Since RD returns a combined measure of how consistent (representative) and distinctive a feature is with respect to a comparison author/authors, no attention is paid to the question, which author used a feature more frequently than the other if the feature is representative for both. Thus, in contrast to the RF information that makes it possible to attribute particular features to authors, features may appear in both lists. Since we are only given the forty to sixty most prominent features for each participant, an exact rankings comparison is not possible in this case. Instead, we also consider the same number of most prominent representative and distinctive features and compare how many items are

Table 9 Dickens' markers, when compared to Collins, according to Tabata's work using RF

Dickens' markers
very, many, upon, being, much, and, so, with, a, such, indeed, air, off, but, would, down, great, there, up, or, were, head, they, into, better, quite, brought, said, returned, rather, good, who, came, having, never, always, ever, replied, boy, where this, sir, well, gone, looking, dear, himself, through, should, too, together, these, like, an, how, though, then, long, going, its
Collins' markers
first, words, only, end, left, moment, room, last, letter, to, enough, back, answer, leave, still, place, since, heard, answered, time, looked, person, mind, on, woman, at, told, she, own, under, just, ask, once, speak, found, passed, her, which, had, me, felt, from, asked, after, can, side, present, turned, life, next, word, new, went, say, over, while, far, london, don't, your, tell, now, before

Table 10 Tabata's Dickens markers, when compared to the larger reference corpus

Positive Dickens' markers
eyes, hands, again, are, these, under, right, yes, up, sir, child, looked, together, here, back, it, at, am, long, quite, day, better, mean, why, turned, where, do, face, new, there, dear, people, they, door, cried, in, you, very, way, man
Negative Dickens' markers
lady, poor, less, of, things, leave, love, not, from, should, can, last, saw, now, next, my, having, began, our, letter, had, I, money, tell, such, to, nothing, person, be, would, those, far, miss, life, called, found, wish, how, must, more, herself, well, did, but, much, make, other, whose, as, own, take, go, no, gave, shall, some, against, wife, since, first, them, word

Table 11 Comparison of highest-rated words under each method for both experiments. Bold printed words indicate a direct correspondence with the other method. Features printed in italic are indirectly shared, namely by the opposing author.

Dickens		**Collins**		**Dickens**		**World**	
RF	**RD**	**RF**	**RD**	**RF**	**RD**	**RF**	**RD**
very	first	**first**	upon	eyes	till	lady	head
many	**upon**	*words*	**first**	hands	for	poor	old
upon	only	**only**	very	again	however	less	looking
being	*left*	*end*	such	are	give	of	returned
much	*words*	*left*	*many*	these	without	things	round
and	*letter*	*moment*	being	under	cannot	leave	down
so	*end*	room	*so*	right	upon	love	door
with	*moment*	*last*	*indeed*	yes	looking	*not*	night
a	*enough*	*letter*	**only**	up	*not*	from	gentleman
such	*answer*	to	*much*	sir	than	should	mr
indeed	*last*	*enough*	*air*	child	but	can	to
air	**such**	back	on	looked	nor	last	*here*
off	**very**	*answer*	a	together	about	saw	through
but	**being**	leave	great	*here*	would	now	face
would	on	still	*and*	back	head	next	its

shared, when the same number of input features is considered, in this case the 300 most frequent ones. Table 11 shows comparisons of the experiments. The number of directly shared items, for instance, items appearing under Dickens under both RF and RD is fairly high—RD shares eighteen words, or ~30% of the sixty most prominent words for Dickens under RF. Considering Collins, the overlap is comparable, namely twenty-one shared items of sixty-six words under RF (~32%). However, what is noticeable is that some of Tabata's 'Dickens features' appear among our 'Collins features', suggesting that they are good separators for the two authors, being more frequent for Dickens, but

more representative for Collins. Regarding the Dickens/reference set comparison, there are two shared items for the forty most prominent words for Dickens under each analysis, while there are twelve out of sixty-two for Dickens' negative words/the reference corpus.

However, if we raise the number of features in the input, using ∼5,000 for the Dickens/Collins comparison, the number of shared items for Dickens falls to four out of sixty and eleven out of sixty-six for Collins. Considering ∼4,000 most frequent words instead of 300 for Dickens/the reference corpus causes a drop to zero out of forty shared words for Dickens and one out of sixty-two for the corpus. The fact that the two methods are similar, given a more limited input is not necessarily surprising, but it indicates that while RF performs better on a few, more-frequent features, this is not true for RD. Comparing the corresponding ARI scores for those 300 input features confirms this; for the two-author experiment, the ARI is also high, but starts dropping relatively quickly on clustering the first 200–250 most prominent features. For the second comparison, the numbers become even less stable, which suggests, that the method struggled more on finding discriminators when only considering the 300 most frequent features.

Thus, the above comparisons indicate that methods are more similar for two-class problems, although this could also be due to the fact that RD might possibly be less suited for mixed set comparisons.

4.4 Comparing to Burrows' Delta

In order to understand to what extent RD is similar or different to other methods extant in the literature, we compare the features emerging from our analysis to those selected (or used) by two other techniques. We begin with a comparison to Burrows' Delta (Burrows, 2002).

From a theoretical point of view, one central difference between the techniques is one of design; Burrows' Delta was intended for authorship attribution, i.e. measuring similarity between a test document and different candidate authors, indicating which author of those considered would be most likely to have authored this particular document. However, RD aims at detecting characteristic stylistic features—thus one question addressed here would be to what extent characteristic stylistic features coincide with those found most discriminating in successful authorship attribution.

Burrows' Delta is an authorship attribution technique used to identify the most likely author for a test document on the most frequent words (1–800 mfw). To perform the test, a corpus of candidate authors is assembled with a couple of documents each, and both the mean and standard deviation for all features are calculated over the complete set of features (words). To compute z-scores for individual authors, for each author and feature, one takes the average standardized frequency over his documents and computes z-scores using mean and standard deviation over the whole corpus. The test document is treated similarly also using the corpus' $\hat{\mu}$ and $\hat{\sigma}$. We then compare the test piece's scores to those of a candidate author and take the mean over the absolute differences to obtain a combined score.

Thus, Delta is defined as 'the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text' (Burrows, 2002). The 'Delta scores' emerging from the analysis quantify the individual comparisons for each author in the main corpus and a specific test piece, where the lowest distance indicates the closest fit. The 'Delta z-scores' refer to z-scores computed over the distribution of Delta scores, e.g. if a value (corresponding to the lowest distance) diverges a lot (from the mean of all differences), it indicates that the author's piece and the test piece are unusually close and that there is no other close competitor (this can be quantified through the z-distribution).

4.4.1 *Delta experiment*

Since the two methods have different aims, there is no direct way of comparing the results. The output of Delta are Delta scores and Delta z-scores corresponding to an aggregation over some number of most frequent words—this does not immediately reveal which words were determining the overall proximity or non-proximity to a test document. To determine what features were central in the analysis, one could examine z-scores of individual features before they are combined into the overall

Table 12 Delta z-scores for candidate authors in corpus w.r.t test text *Nicolas Nickleby*, indicating that Dickens is not notably closer to the test document than the other candidates

Author	Delta z-score
Dickens	−0.65
Eliot	−0.53
C. Brontë	−0.50
Gaskell	−0.50
Thackeray	−0.48
Collins	−0.48
Trollope	−0.48
Smollett	−0.41
Austen	−0.41
Sterne	−0.39
Swift	−0.38
Fielding	−0.38
Richardson	−0.34
Defoe	−0.33
E. Brontë	1.98
Goldsmith	2.13
A. Brontë	2.15

Table 13 Rank correlation of different numbers of features based on Delta and RD; where a high negative correlation would be indicative of a strong similarity between the methods

Number of features	Spearman's ρ
800	–0.17
700	–0.16
600	–0.16
500	–0.12
400	–0.09
300	–0.04
200	–0.02
100	–0.01
50	0.11
20	–0.28
10	–0.13
5	0.80
2	–1.00

Delta score. For instance, important features for Dickens should show low absolute differences between z-scores of Dickens' set and one of his documents as a test document.

In the following experiment, we consider a classic Delta analysis as well as one that allows for a comparison to characteristic features emerging from applying RD to the same data. The data set used for the analysis is the same as the one used in Section 4.2. More specifically, there are twenty-four texts by Dickens and fifty-five by sixteen other authors. Although this would be a suitably balanced set for RD, it is less well suited for applying Delta due to the fact that Dickens is dominating as a single author. For this reason, we reduce Dickens' set in order to prevent his style from dominating the mean and standard deviation over the entire corpus—which are crucial parameters for Delta. We randomly extract eight documents for Dickens and take the remainder as test pieces. The data was preprocessed as described in Section 3. For the final input, we retain the 800 most frequent features.

First, considering a classic Delta analysis of the data, the Delta scores reveal that in all sixteen cases, Dickens is rated closest to his own document. Considering the distributions of Delta over all authors, namely Delta z-scores, it seems that under Delta Dickens' documents are not extraordinarily similar to one another based on these test pieces and when compared to the other candidate authors (A typical result is shown in Table 12).

4.4.2 *Feature comparison*

In order to compare the two methods, we use the same training data (sixty-three authors on 800 features) to compute representative and distinctive features (for Delta, we consider the feature values corresponding to Table 12). To examine similarities in feature importance, we can compare the rankings of the features under the two methods. For Delta, low values indicate greater importance, while in terms of RD, higher values would be more desirable. We correlate the rankings for all 800 features under each method using *Spearman's* ρ, which is bounded by $[-1,1]$. Thus, for a strong correlation in the present case, we would expect a large negative correlation. Correlating all the rankings over all 800 features returns a weak negative value: −0.17, however, among those 800, there might be less-accurate ones, so it remains to test higher-rated features' correlations. For this purpose, we reorder the features according to the highest representative and distinctive features and try different levels of highest values, shown in Table 13. The correlation between the

number of features considered and the correlation between methods is −0.67, the mean of this over all sixteen test pieces is −0.49, with correlations ranging from −0.1 to −0.7, which does not indicate a very stable relationship. But this does indicate that it is beneficial to include a larger number of features (words). Thus, the degree of correlation seems to be subject to the particular test document, as well as the composition of test and training corpus.

Further, we can compare the number of top features shared between the methods. Among the first approximately twenty to thirty most important features, methods share only one term, namely 'hardly'. Among the first 100 words, there are twenty-two shared ones: 'more', 'nothing', 'without', 'however', 'old', 'hardly', 'she', 'return', 'for', 'entered', 'stay', 'about', 'future', 'but', 'conduct', 'away', 'pleased', 'immediately', 'entirely', 'cold', 'be', and 'than'. Considering the first 200 most important ones yields sixty-three shared features; the first 300 raises it to 132 common features.

The above comparison showed that there might not be a very strong or even consistent correlation between features emerging as important from the two methods. Delta scores (per feature) and RD_f scores correlate only weakly, from which we conclude that they are genuinely different. However, since they were designed for different purposes, any comparison between them is unlikely to be ideal. In our case, Delta requires that one includes fewer documents by Dickens in the main corpus, while more documents would be better for RD to estimate Representativeness more reliably. Generally, features that are consistent for a particular author in terms of being avoided or preferred with respect to the main corpus, are likely to emerge under both methods, provided the chosen test piece is also following this regular pattern.

4.5 Comparing to Hoover's CoV Tuning

For the comparison between the CoV Tuning method (Hoover, 2014) and RD, we again consider the Dickens/Collins data set.

The CoV Tuning method was introduced to 'identify words used fairly frequently and in many texts but with widely varying frequencies'. For this purpose, one considers a two-/multi-author text corpus and computes the Coefficient of Variance over the complete sample (for each feature f separately) by dividing the standard deviation σ_f by the mean μ_f (the computations are on the basis of relative frequencies). The resulting scores are then multiplied by 100 to express them as percentages. However, Hoover notes that high CoVs are also awarded to features that are rare or only occur in a small number of texts, which necessitates choosing items that occur in a large number of texts. According to David Hoover (email communication), there do not yet exist clear guidelines for choosing the number of documents a term has to appear in, so this is done here heuristically as well.

4.5.1 *CoV tuning experiment*

Since the methods operate on different levels of the data set, i.e. CoV Tuning being computed on the basis of the whole corpus and RD requiring division of authors into sets, there is unlikely to be an ideal experimental design for comparison. Similar to the previous experiment, there are different aspects one may consider to gain some intuition about the similarities and differences between the two techniques. To arrive at a good estimation for thresholds of input features, we analyze accuracy in clustering documents for the highest features under the CoV Tuning method. Further, we examine similarities with respect to the features chosen by the CoV as highest and look at the CoV and RD_f score correlations for these features. Finally, we consider highly rated words shared by both methods, when RD is applied as usual.

4.5.2 *Clustering with the CoV*

In order to restrict the number of input features, different thresholds were explored, but only a very high threshold of 'appearance in at least 98% of the documents' proved effective in terms of clustering (practically, this included features appearing in all documents). This reduced the data to 1,063 input features. Table 14 shows the results for clustering different levels of top features for the CoV. The distance matrix was computed using the 'Manhattan' distance and clustering was done using 'complete link'. The clustering result is evaluated using the ARI. The results indicate, that in this case, at least

Table 14 CoV Tuning's accuracy in clustering on the Dickens/Collins set, shown using different numbers of highest input features

Number of features	ARI
300	0
350	0.69
400	0.84
500	0.84
550	0.84
600	0.76
650	0.76
700	0.84
800	0.84
850	0

350 features are required, and clustering results are highest on 400–800 features.

4.5.3 *Comparing CoV tuning and RD*

In order to investigate correlations between the two methods, we consider the highest features emerging under CoV Tuning with respect to clustering and consider the exact same features ordered by their RD_f scores. A high correlation in terms of rank would be marked by a high *Spearman's* ρ, close to 1. Table 15 shows selected levels of the ranking correlations of CoV and RD_f scores for both Dickens and Collins. Occasionally, there are stronger correlations for Collins' scores and the CoV, but since these are also negative, it seems rather erratic. The correlation between the number of features considered and the correlation between methods is 0.54 for Dickens and 0.73 for Collins, which indicates that the level is likely to be relevant here (the overall correlations were computed on a stepwise version of the data, e.g. for 1,000 levels, there were ~1,000 correspondences). We interpret the low correlation to indicate that CoV and RD are genuinely different concepts.

4.5.4 *Shared feature lists*

As a final exercise, we look into size and type of features identified by the two methods where RD is computed on the entire feature input of ~5,000 features. Since the method is computed with respect to particular author samples, less frequent, but consistent features are considered likewise. Thus, for each method, we order features according to prominence and consider the overlap at different levels of the ranked list.

Table 15 Correlation of rankings on various levels of top features according to the features selected for the CoV

Number of features	Spearman's ρ	
	Dickens	Collins
1,000	0.07	0.13
900	0.09	0.10
800	0.09	0.09
700	0.10	0.07
600	0.12	0.02
500	0.11	–0.03
400	0.15	–0.03
300	0.09	–0.08
200	0.01	–0.19
100	–0.07	–0.25
50	–0.08	–0.38
40	–0.06	–0.36
30	0.04	–0.21
20	–0.12	–0.25
10	–0.04	0.41
5	0.10	1.00

Table 16 Number of shared items at different levels of prominence, including the top features—for RD for both all original input features before 'Tuning' and only using the features input to CoV computations

Number of features	*Input*		
	5,000 mfw		1,063 CoV
	Dickens	Collins	Dickens
500	117	132	241
400	86	86	152
300	57	59	101
200	34	37	52
150	21	23	31
100	8	11	12
90	5	7	6
80	5	5	5
70	3	4	4
50	2	2	2
40	2	1	2
30	0	0	0

Table 16 shows the number of shared items at different steps. When considering both Dickens and Collins (for all 5,000 features as input), the overlap

with the features selected by the CoV is not considerable—the top 100 features only yield eight to eleven shared items, but which incidentally include 'upon' and 'letter', which have previously been identified as Dickens and Collins markers (Tabata, 2012). Further, we compare the features chosen by CoV and RD (for Dickens) on the exact same input of 1,063 features appearing in all documents. The overlap of highest-ranked features is greater after the first 100 words, but less than one might expect on the same input, if the methods were choosing features in a similar fashion.

In terms of a general comparison, we note that CoV Tuning requires virtually no computation time compared to the expensive pairwise comparisons of documents needed for RD.

Disregarding any particular author in the set (unsupervised approach), as it is done in CoV Tuning, potentially offers more possibilities for evaluation than a supervised technique, where accuracy of selected features can only be heuristically evaluated for instance, by clustering. The fact that CoV Tuning is successful at all, considering it operates only by measuring variability of frequent features is impressive—however, this potentially indicates a different application area than RD, where the focus is on author-dependent consistency of usage regardless of exact frequency strata. There is an overlap, nevertheless, if only at a theoretical level, as items appearing in most documents as well as being highly variable might be more likely to vary between than within authors.

5 Conclusion

This work has introduced RD, a simple statistical measure to identify features that an author uses consistently and in a way that distinguishes him/her from others. The technique requires a substantial number of documents of each author (in order to gauge consistency), and its performance wanes when one set is less homogenous. Different comparisons to other techniques applied in the domain, both well established and recently introduced ones, indicate more differences than similarities to RD. Its ability to analyze both frequent as well as less-frequent features renders it a powerful and promising technique for stylometric analysis in authorship.

5.1 Future considerations

We should like to be able to characterize the extent to which one can consider a feature score high or low in an absolute sense as opposed to merely high or low with respect to the other features for a particular author. For instance, there are authors, such as Jane Austen, who are rather consistent in vocabulary use throughout their different works and who might thus be more likely to end up with higher representative scores than authors displaying less consistency, such as for instance Mark Twain, who is seen to be more volatile. Future work might therefore include exploring the properties of high and low RD_f scores in order to be able to generalize about the degree to which an author is consistent over his works and different from others.

Our goal in this article was to suggest an emphasis in stylometry on features whose frequency distributions might be regarded as fairly characteristic for a given author as opposed to those that serve to discriminate the author from others. Our comparisons have indicated that these two characterizations may be very different. As stylometry evolves to encompass syntactic features, which we suspect will be less numerous than the very large vocabularies of authors, the shift in emphasis may become more important.

References

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1): 5–32.

Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3): 267–87.

Burrows, J. (2005). Who wrote Shamela? Verifying the authorship of a parodic text. *Literary and Linguistic Computing*, 20(4): 437–50.

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

Craig, H. (1999). Authorial attribution and computational stylistics: if you can tell authors apart, have you

learned anything about them? *Literary and Linguistic Computing*, 14(1): 103–13.

Craig, H. and Drew, J. (2011). Did Dickens write "Temperate Temperance"?: (an attempt to identify authorship of an anonymous article in all the year round). *Victorian Periodicals Review*, 44(3): 267–90.

Hoover, D. (2007). Corpus stylistics, stylometry, and the styles of Henry James. *Style*, 41(2).

Hoover, D. (2014). Tuning the word frequency list. *Digital Humanities 2014: Conference Abstracts*, Université de Lausanne, pp. 200–2

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1): 193–218.

Mahlberg, M. (2007). Clusters, key clusters and local textual functions in Dickens. *Corpora*, 2(1): 1–31.

Manning, C. D., Raghavan, P. and Schütze, H. (2008). *Introduction to Information Retrieval*, Vol. 1. Cambridge: Cambridge University Press.

Oakes, M. P. (2014). *Literary Detective Work on the Computer*, Vol. 12. Amsterdam: John Benjamins Publishing Company.

Prokić, J., Çöltekin, Ç. and Nerbonne, J. (2012). Detecting shibboleths. *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, EACL Avignon, 2012. Association for Computational Linguistics, pp. 72–80.

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E. and Inches, G. (2013). Overview of the author profiling task at pan 2013. *Notebook Papers of CLEF*, pp. 23–6. www.uni-weimar.de/medien/webis/research/events/pan-13/pan13-web/

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3): 538–56.

Tabata, T. (2012). Approaching Dickens' style through Random Forests. *Digital Humanities: Conference Abstracts*, Hamburg: University of Hamburg, pp. 388–91.

Wieling, M. and Nerbonne, J. (2011). Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language*, 25(3): 700–15.

Notes

1 It has been suggested that work in author profiling might be relevant to the task of finding typical features, and this is indeed similar, but the focus of profiling is rather on distinguishing groups of authors, e.g. by age or sex. See Rangel *et al.* (2013) and references there.

2 All computations for this article, including Representativeness and Distinctiveness were implemented using the statistical language R (R Core Team, 2014), using packages, such as 'cluster', 'stats', and 'mclust'.

3 We would like to thank Tomoji Tabata for making his data set available to us.