Towards Distinctive and Typical Style Features in Authorship

Carmen Klaussner[*] KLAUSSNC@TCD.IE
Çağri Çöltekin[**] C.COLTEKIN@RUG.NL
John Nerbonne[**] J.NERBONNE@RUG.NL

[*] *Trinity College Dublin, Ireland*

[**] *University of Groningen, The Netherlands*

Abstract

Detection of stylistic elements in authorship studies is hampered by the lack of a gold standard that would otherwise enable us to clearly evaluate our findings. In absence thereof, one generally resorts to choosing items for which an author shows a characteristic usage compared with other writers. In this line of work, we present both a measure for determining characteristic elements of an author that he uses consistently over different works by examining different types of features, both lexical and syntactic ones. For evaluation, we test the separation ability of the selected features by clustering the data set on their basis.

We apply both feature selection and evaluation in two different studies of authorship. In the first, we compare Charles Dickens and Wilkie Collins, while the second one is contrasting the styles of Henry James and Mark Twain. Testing separation ability in clustering on highly representative and distinctive features returns results very close to the ideal clustering result.

1. Introduction

The concept of *style* in studies of authorship refers to the something like the feel of a piece of text, which might be less tangible than style in other disciplines, such as art or music, where we use *language* to express aspects of a painting or a piece of music, whereas for style in writing, we have to use the same tools that were used to create it.

While traditional studies of authorship involve individual persona judging upon what constitutes the style of an author, non-traditional studies employ statistical techniques to discover the style of an author. In either scenario, the development and investigation of suitable methods of detection is hampered by the lack of a gold standard, that would be able to indicate the method's closeness to returning true stylistic elements of an author. In absence thereof, the weight lies primarily on the method's theoretical appropriateness in the way it selects stylistic markers, which results in the need to understand our methods as well as their implications.

Consistency of stylistic markers has been somewhat of a central theme to studies of authorship, such as in early studies (Mosteller and Wallace 2008), which also emphasised collecting larger number of markers to increase reliability of the result. This has been continues in more recent studies as in the development of Burrow's Delta (Burrows 2002) and Zeta (Burrows 2007) measures that build on features representative for an author over a given number of his texts, while distinguishing him from another/other author(s).

In this work, we consider *Representativeness & Distinctiveness* (Prokić et al. 2012), a technique originating in dialectometry to select an author's consistent features that are at the same time also distinctive with respect to an opposing author's sample. Further, we propose a heuristic method for evaluation of those selected features intended to measure how well they are able to separate the data set into the correct groupings. As part of the analysis, we include both lexical and syntactic ones, such as simple word uni-grams, but also Part-of-Speech (POS) bi-grams/tri-grams. The data consists of two separate authorship sets, where the first consists of texts by two British authors, Charles Dickens

and Wilkie Collins and the second comprises writings by the two American authors, Henry James and Mark Twain. Related work in stylometry includes a study of Dickens' style compared to both Collins and a larger reference set comprising texts from the 18th and 19th century (Tabata 2012). Further, there has been a corpus stylistics study (Mahlberg 2007), aiming to extract Dickens-specific key word clusters (sequences of words), that can be interpreted as pointers to more general functions, such as *Body Part* clusters pertaining to Dickens' particular affinity for using body parts for individualisation of characters, e.g. "his hands in his pockets".

Thus, in section 2, we describe the two data sets, while we introduce *Representativeness & Distinctiveness* in section 3. Further, section 3.2 introduces and illustrates our proposed method of evaluation. Finally, in section 4, we describe our experiments and results with respect to the two data sets and section 5 closes the discussion.

2. The Data Sets

For this study, we built two different author sets, one comparing Charles Dickens and contemporary writer Wilkie Collins and the other one opposing Henry James to fellow American writer Mark Twain. All data was obtained from *Project Gutenberg*[1] and the *Internet Archive*[2].

2.1 Dickens vs. Collins

Table **??** and table 1 show Dickens' and Collins' data set respectively, both comprising 27 documents. Dickens' data set does not only contain documents by himself, but three books, namely, *A Budget of Christmas Tales*, *A House to Let* and *No Thoroughfare* are collaborations, where Dickens seems to have been main author. In particular, the last two also include Wilkie Collins as author, where *No Thoroughfare* was written by only Dickens and Collins. These were included, since they might be interesting with respect to stylistic properties. If both authors persist in terms of style, these should be somewhat more difficult to classify than those written by only one of them.

2.2 James vs. Twain

While Dickens and Collins seemed to have been close enough to collaborate on work, this is a highly unlikely scenario for Henry James and Mark Twain. Although both being American writers close in age, they did not seemed to have approved of each other as artists (Canby 1951). It is interesting to investigate to what extent this mutual dislike and disapprobation of each other 's work manifests itself in their writings and what elements separate them. Table 2 and table 2 show James' and Twain's data sets, where James is represented with 25 and Twain with 21 works.

1. http://www.gutenberg.org/
2. https://archive.org/

No.	Author	Texts	Abbr.
1	Dickens	Bleak House	D1023
2	Dickens	Great Expectations	D1400
3	Dickens	Little Dorrit	D963
4	Dickens	David Copperfield	D766
5	Dickens	A Christmas Carol	D19337
6	Dickens	Life And Adventures Of Martin Chuzzlewit	D968
7	Dickens	The Mystery of Edwin Drood	D564
8	Dickens	A Tale of Two Cities	D98
9	Dickens	Master Humphrey's Clock	D588
10	Dickens	The Battle of Life: A Love Story	D40723
11	Dickens	Life And Adventures Of Nicholas Nickleby	D967
12	Dickens	Barnaby Rudge	D917
13	Dickens	Sketches of Young Couples	D916
14	Dickens	The Uncommercial Traveller	D914
15	Dickens	Our Mutual Friend	D883
16	Dickens	Pictures From Italy	D650
17	Dickens	Sketches by Boz	D882
18	Dickens	A Child's History of England	D699
19	Dickens	Reprinted Pieces	D872
20	Dickens	Dombey and Son	D821
21	Dickens	Oliver Twist	D730
22	Dickens	The Old Curiosity Shop	D700
23	Dickens	American Notes	D675
24	Dickens	The Pickwick Papers	D580
25	Dickens (et al.)	A Budget of Christmas Tales	Dal28198
26	Dickens (et al.)	A House to Let	Dal2324
27	Dickens (/Collins)	No Thoroughfare	DC1423

Table 1: Collins' data set.

No.	Author	Texts	Abbr.
1	Collins	After Dark	C1626
2	Collins	Antonina	C3606
3	Collins	Armadale	C1895
4	Collins	Man and Wife	C1586
5	Collins	Little Novels	C1630
6	Collins	Jezebel's Daughter	C3633
7	Collins	I Say No	C1629
8	Collins	Hide and Seek	C7893
9	Collins	Basil	C4605
10	Collins	A Rogue's Life	C1588
11	Collins	The Woman in White	C583
12	Collins	The Two Destinies	C1624
13	Collins	The Queen of Hearts	C1917
14	Collins	The New Magdalen	C1623
15	Collins	The Moonstone	C155
16	Collins	The Legacy of Cain	C1975
17	Collins	The Law and the Lady	C1622
18	Collins	The Haunted Hotel: A Mystery of Modern Venice	C170
19	Collins	The Fallen Leaves	C7894
20	Collins	The Evil Genius	C1627
21	Collins	No Name	C1438
22	Collins	Poor Miss Finch	C3632
23	Collins	Rambles Beyond Railways	C28367
24	Collins	The Black Robe	C1587
25	Collins	Miss or Mrs.?	C1621
26	Collins	My Lady's Money	C1628
27	Collins	The Dead Alive	C7891

Table 2: Twain's data set.

No.	Author	Texts	Abbr.
1	Twain	Innocents Abroad	T-ia
2	Twain	The Gilded Age: A Tale of Today	T-tgaatot
3	Twain	Sketches New and Old	T-snao
4	Twain	The Adventures of Tom Sawyer	T-taots
5	Twain	A Tramp Abroad	T-ata
6	Twain	Roughing It	T-ri
7	Twain	The Prince and the Pauper	T-tpatp
8	Twain	Life on the Mississippi	T-lotm
9	Twain	The Adventures of Huckleberry Finn	T-taohf
10	Twain	A Connecticut Yankee in King Arthur's Court	T-acyikac
11	Twain	The American Claimant	T-tac
12	Twain	The Tragedy of Pudd'nhead Wilson	T-ttopw
13	Twain	Tom Sawyer Abroad	T-tsa
14	Twain	Tom Sawyer Detective	T-tsd
15	Twain	Personal Recollections of Joan Arc	T-proja
16	Twain	Following the Equator: A Journey Around the World	T-fteajatw
17	Twain	Those Extraordinary Twins	T-tet
18	Twain	A Double Barrelled Detective Story	T-adbds
19	Twain	Christian Science	T-cs
20	Twain	Chapters from My Autobiography	T-cfma
21	Twain	The Mysterious Stranger	T-tms

Table 3: James' data set.

No.	Author	Texts	Abbr.
1	James	The American	J-ta
2	James	Watch and Ward	J-waw
3	James	The Europeans	J-te
4	James	Confidence	J-c
5	James	Washington Square	J-ws
6	James	Portrait of a Lady	J-poal
7	James	Roderick Hudson	J-rh
8	James	The Bostonians	J-tb
9	James	Princess Casamassima	J-pc
10	James	The Reverberator	J-tr
11	James	The Aspern Papers	J-tap
12	James	The Tragic Muse	J-ttm
13	James	The Other House	J-toh
14	James	What Maisie Knew	J-wmk
15	James	The Spoils of Poynton	J-tsop
16	James	Turn of the Screw	J-tots
17	James	The Awkward Age	J-taa
18	James	The Sacred Fount	J-tsf
19	James	The Wings of the Dove	J-twotd
20	James	The Golden Bowl	J-tgb
21	James	The Ambassadors	J-tamb
22	James	The Outcry	J-to
23	James	The Ivory Tower (unfinished)	J-tit
24	James	The Sense of the Past (unfinished)	J-tsotp
25	James	In the Cage	J-itc

3. Representativeness and Distinctiveness for Stylometry

The statistical technique of Representativeness and Distinctiveness was originated in the realm of dialectometry, where it has been shown to detect lexical items able to distinguish different dialectical areas (Prokić et al. 2012). In the context of stylometry, the method can be employed to detect elements for which an author is consistent throughout his own works while also separating him from others. Considering, for instance, a comparison between Dickens and fellow writer Collins on word features using a couple of novels of each writer, one first determines Dickens' representative terms, i.e. those words which he uses consistently either frequently or infrequently over his works. In order to arrive at a combined measure, one then favours those representative terms of Dickens that Collins uses either inconsistently or consistently but with a different frequency over his novels. The remaining group of words are considered to be Dickens' representative and distinctive terms when compared with Collins.

Thus, Representativeness and Distinctiveness bears similarities with both Burrow's Delta (Burrows 2002) and Zeta (Burrows 2007) in so far as favouring consistent terms that are irregular in the opposing author's set. Additionally, it is also similar to Zeta in being dependent on the other set for the selection of distinctive terms out of the representative ones. Formally the method of Representativeness and Distinctiveness is defined as follows:

Representativeness of a feature f for document set D is defined in eq. 1.

$$\overline{d_f^D} = \frac{2}{|D|^2 - |D|} \sum_{d,d' \in D, d \neq d'} d_f(d, d') \tag{1}$$

The Distinctiveness measure for comparing to outside documents corresponds to eq. 2.

$$\overline{d_f^{D'}} = \frac{1}{|D|(|DS| - |D|)} \sum_{d \in D, d' \notin D} d_f(d, d') \tag{2}$$

The distance d_f between document d and d' with respect to feature f, is set as the absolute difference between the logarithm of the relative frequency of their respective original term frequencies (eq. 3). Relative frequency is given preference here to normalise over document size, while taking the logarithm lessens the effect of rather high frequencies.

$$d_f(d, d') = |log(relFreq(f) - log(relFreq(f')| \tag{3}$$

$\overline{d_f^{D'}}$ and $\overline{d_f^D}$ are standardised by using all distance values calculated for feature f to yield the degree of representativeness and distinctiveness for feature f in D with respect to DS as defined in eq. 4.

$$\frac{\overline{d_f^{D'}} - \overline{d_f}}{sd(d_f)} - \frac{\overline{d_f^D} - \overline{d_f}}{sd(d_f)} \tag{4}$$

3.1 Determining Stylistic Author Profiles

When applying Representativeness & Distinctiveness for building author profiles, one needs to consider that there is an ambiguity with respect to a term being distinctive from one author to another. We consider the previous case, where we are trying to find features that separate Dickens and Collins. Initially, we identify Collins' representative terms, which is only based on comparing his own documents for different features. Once, this is completed, we identify those terms that separate Collins' set well from Dickens' set of documents. Finally, we choose the highest representative ones that also succeed in being discriminative between Collins and Dickens.

However, this analysis is directional in the sense that the degree of representativeness of a certain feature could be different for the opposing set, i.e. there are two different scenarios for a feature being *different* in the opposing set.

Case 1. The term t_i is consistent in Collins' set C with a low frequency, while the same term t_i is consistent in Dickens' set D with a high frequency. Thus, the term is representative and distinctive for both sets, even though we did not consider the *Representativeness* for set D. Obviously, the converse could also be true: a consistently high frequency for set C and a consistently low frequency for the set D. This first case does not produce any issues for measuring similarity, since on the basis of these features there is is reliable similarity within sets and accentuated differences between the sets.

Case 2. The second possibility is the one that may cause issues. Assuming a representative and distinctive term for set C, with a frequency either high or low. However, the same term is not representative for set D and values may fluctuate from high to low. Although this term is not representative for D, it is distinctive from C to D, because it is constant in C while not being so in D. Clustering the data set on the basis of these terms may create noise, since it will not show similarities for documents within D and may have occasional rather similar values to the ones in C that rate it closer to documents in C.

In order to identify elements representative and distinctive for both sets, first, the analysis has to be carried out for both sets and having identified two separate author profiles, we select those features that are shared by both.

Figure 1: Representativeness

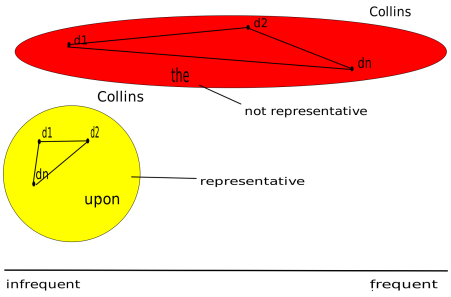
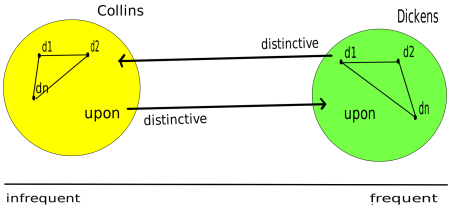
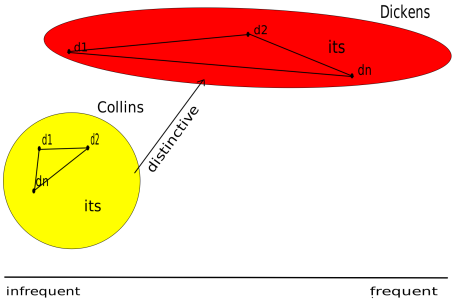


Figure 2: Distinctiveness

Case 1



Case 2



3.2 Evaluation of Distinctive Markers

Evaluation in studies of authorship can only be done heuristically by identifying desirable properties of stylistic markers and determining means of measuring this property. One of these desirable characteristics could be considered as stylistic marker's ability to separate the author's set in question. For instance, having identified a set of discriminatory markers for Dickens and Collins using a particular method, we evaluate that method by testing whether those markers are indeed able to cluster the document space appropriately into the two groups of documents. Thus, discrimination ability of a set of distinctive markers is determined by the overall similarity grouping of documents according to those markers. Poor discriminators will result in poor clustering, where authors are not separated well into two groups, while good discriminators should be able to divide documents of different origin clearly into two different groups. This can be quantified by evaluating the clustering result using the *Adjusted Rand Index* (Hubert and Arabie 1985), which

compares the current clustering result to the ideal result by pairwise comparison of the groups. The returned index is bounded by [-1,1], with 0 being the expected value and 1 the highest positive correlation between two different clustering versions.

4. Experiments

As part of the experiments, we are looking at both data sets for different feature types, in this case simple word uni-grams, POS bi-grams and tri-grams. The best features for each individual author are selected by computing a threshold based on the representative and distinctive scores of all features within. Thus, for an authors full list of features, one first computes the mean μ_f over all representative and distinctive values for all features and multiplies this by a fixed value, e.g. 1.8 that is dependent on the feature input size and individual requirements. In this case, only features with a score over $\mu_f \times 1.8$ would be kept.

In section 4.1, we look at distinctive features of Dickens compared to Collins and section 4.2 opposes contemporaries James and Twain.

4.1 Dickens vs. Collins

The Dickens and Collins data set individually both contained 27 novels, where three in Dickens' set were collaborations with other author and one only with Collins. The uni-gram input matrix was a 54-documents x 5000 features matrix, where the features were the 5000 most frequent features. Similarly, for POS tri-grams, there was a 54 x 5000 input matrix and for POS bi-grams, it was 54 x 1100. The author profiles for word uni-grams and POS tri-grams were chosen by using a threshold of $\mu_f \times 1.9$ and for POS bi-grams by $\mu_f \times 1.7$.

Table 4: Dickens' and Collins' highest rated features

No	**Word uni-grams**		**POS bi-grams**		**POS trigrams**	
	Dickens	Collins	Dickens	Collins	Dickens	Collins
1	upon	upon	RP.CC	NN.TO	CC.NN.CC	IN.VBG.PP
2	scarcely	discovered	IN.IN	RB.TO	NN.CC.RB	TO.DT.NN
3	discovered	produced	NN.TO	RP.CC	RB.IN.IN	IN.VBG.DT
4	many	interests	VBG.RP	TO.PP	CC.VBG.RP	PP..JJ.JJ
5	and	left	CC.RB	VBP.VBD	VBN.RP.CC	DT.NN.TO
6	left	many	CC.VBG	NN.NN	RB.MD.VB	VBD.IN.VBG
7	very	useless	JJ.CC	NNS.NN	RB.TO.PP	VBN.IN.VBG
8	but	attempt	NN.CC	RB.NN	CC.VBG.IN	DT.NN.PP
9	much	motive	CC.IN	NP.NN	DT.NN.PP	NN.NN.NN
10	beside	only	DT.CC	RB.JJ	RB.JJ.CC	NN.WDT.PP
11	though	risk	RB.CC	VBZ.VBD	JJ.NN.CC	TO.PP.NN
12	down	resolution	RB.TO	CC.WRB	IN.IN.DT	TO.VB.PP
13	produced	words	VBG.JJ	CC.VBG	CC.VBG.CC	VBD.TO.PP
14	several	hesitated	IN.CC	VBG.RP	RB.RB.CC	RB.TO.PP
15	lest	very	CC.NN	VBN.PP	VBG.RP.CC	PP..NNS.PP
16	failed	scarcely	CC.WRB	NNS.VBP	NNS.CC.RB	VBD.NP.NNS
17	control	future	VBG.CC	VBN.TO	DT.NN.TO	PP.NN.POS
18	great	return	RP.NNS	RB.MD	CC.JJ.CC	PP.DT.NN
19	such	first	PDT.DT	NP.NNS	RB.CC.VBG	DT.NN.PP
20	indeed	failed	CC.JJ	VB.PP	CC.RB.RB	VBD.TO.DT

Table 4 shows the highest rated features according to Representativeness and Distinctiveness that are best in discrimination between the two authors. The values for Dickens' features were

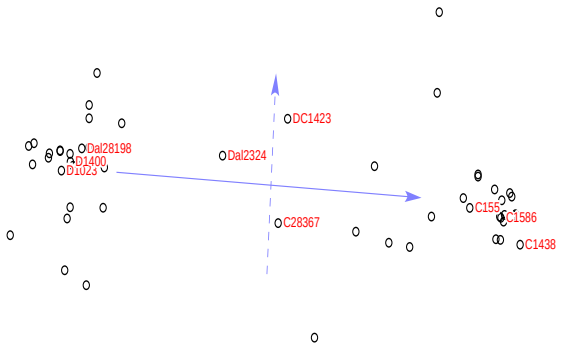


Figure 3: MDS plot on Dickens and Collins highest shared word uni-grams corresponding to iteration 27.

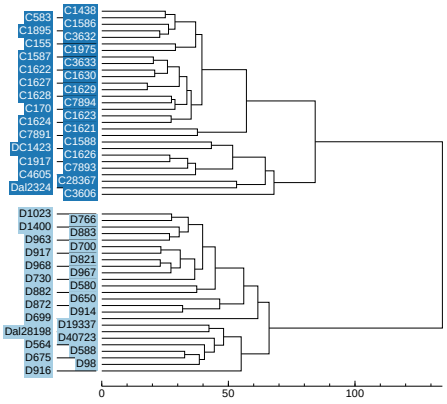


Figure 4: Dendogram on Dickens and Collins highest shared word uni-grams corresponding to iteration 27

calculated by comparing to Collins and vice versa. Column two and three show the top 20 uni-gram features for the two authors separately, while column four and five show POS bi-grams. For the word uni-grams, 8 of the 20 terms appear in both lists, so this corresponds to Case 1 of Distinctiveness, where they are consistent for both authors, but have very different frequency distributions, here UPON, MANY, LEFT, DISCOVERED appear among the top 6 in both lists. For the POS bi-grams, one can observe that 6 features, namely NN.TO, RB.TO, RP.CC, CC.WRB, CC.VBG, VBG.RP are shared by both authors over the first 20 highest items. Noticeable is the frequent appearance of the CC and VBG tag, corresponding to 'coordinating conjunction' and 'Verb in gerund or past particle' respectively. For the POS tri-grams, the list of shared items decreases and only three remain common to both authors for the first 20: DT.NN.TO, DT.NN.PP, RB.TO.PP. Notable here are sequences involving possessive pronoun (PP) and TO (to) after a noun phrase.

The fact that more items are shared for word uni-grams is reflected in the evaluation using clustering, shown in table 6. For clustering we use the features shared by both authors by intersecting both profiles. The validation method used to create individual subsets of the total data set is *Leave-one-out* cross-validation. The results for the Adjusted Rand Index for word uni-grams are consistently high with almost all values being around 93%. In comparison, POS bi-grams fare less well with an average of ... POS tri-grams are slightly better, with an average of ... This might be explained by considering that POS bi-grams are a more general construction than POS tri-grams (there are also less of them) and they might be worse in discriminating between authors of similar backgrounds than tri-grams or even word uni-grams.

Figure 3 and figure 4 show the MDS plot and dendrogram corresponding to the shared features of iteration 27. The dendrogram is computed using the complete link similarity measure.[3] The two author groups seem to form roughly to clusters. Interestingly, the combined Dickens-Collins piece, *DC1423* is (mis)classified as one of Collins' set. One of the other collaborations *Dal2324* is also approaching the border, although from the correct side. The other collaboration, *Dal28198* lies at the centre of Dickens' cluster. Another Collins' piece, *C28367* corresponding to *Rambles Beyond Railways* has been noted as suspicious in the previous study of Dickens/Collins (Tabata 2012). The dendrogram confirms this analysis, showing two distinct clusters, with *Dal28198* and *DC1423* being allocated to Collins' cluster.

3. All MDS plots and dendrograms were created using Gabmap (Nerbonne et al. 2011): http://www.gabmap.nl/.

4.2 James vs. Twain

The James vs. Twain comparison contained 46 documents, where 25 belong to Henry James and 21 to Mark Twain. As before, we test three different feature types, word uni-grams, POS bi-grams and POS tri-grams. For POS bi-grams, there was an 46 x 1100 input matrix and for the two other feature types, POS tri-grams and word uni-grams, it was an 46 x 5000 input matrix. The author profiles for POS tri-grams were chosen by using a threshold of $\mu_f \times 1.9$ and for POS bi-grams and word uni-grams a threshold of $\mu_f \times 1.7$ was used.

Table 5: James' and Twain's highest rated features

No	**Word unigrams**		**POS bigrams**		**POS trigrams**	
	James	Twain	James	Twain	James	Twain
1	companion	companion	NN.CC	NN.CC	DT.NN.CC	CC.VB.DT
2	around	around	CC.VB	CC.VB	NN.IN.WDT	DT.NN.CC
3	and	and	CC.DT	CC.DT	CC.VB.DT	NN.CC.VB
4	conscious	along	PP..RB	IN.WDT	CC.VBD.RB	NNS.CC.RB
5	million	body	CC.JJ	NNS.CC	IN.WDT.PP	NNS.CC.VB
6	spite	might	IN.WDT	RB.TO	NNS.RB.CC	VB.DT.NNS
7	two	declared	RB.CC	CC.NN	TO.PP.IN	NNS.RB.CC
8	might	everybody	NN.RBR	CC.VBD	NN.RB.CC	NN.RB.CC
9	maybe	spite	CC.CD	CC.JJ	VBN.TO.PP	CC.VB.IN
10	rather	chapter	TO.WDT	PP..RB	NN.TO.WDT	CC.VBD.DT
11	pretend	conscious	RP.DT	VB.DT	NN.CC.VB	NN.CC.NNS
12	least	least	CC.VBD	DT.NNS	IN.WDT.PP.	CC.DT.NNS
13	everybody	charming	DT.NNS	CC.NNS	NNS.CC.RB	IN.WDT.PP
14	along	two	NNS.CC	TO.WDT	TO.VB.PP	DT.NNS.CC
15	drunk	having	RBR.RB	CC.RB	PP..RB.JJ	NN.IN.WDT
16	simply	children	NPS.NP	RBR.RB	DT.NNS.CC	RB.TO.VB
17	companions	thunder	WDT.PP.	CC.VBP	RB.VBN.PP	TO.PP.IN
18	her	twelve	WP.PP	RB.CC	RB.DT.NNS	RB.DT.NNS
19	sense	crippled	VBN.PP	PP.IN	NNS.CC.VB	RB.CC.VBD
20	declared	procession	RBR.JJ	CC.CD	CC.DT.NNS	CC.RB.DT

Table 5 shows the 20 highest rated (in terms of Representativeness/Distinctiveness) features for all three feature types. Each list of top author profile features is distinctive with respect to the other author in the comparison. An closer look at common features of Twain and James shows that they share 11 out of 20 uni-grams: COMPANION, AROUND, AND, ALONG, MIGHT, DECLARED, EVERYBODY, SPITE, CONSCIOUS, LEAST, TWO. Of POS bi-grams, 13 out of 20 are common to both authors: NN.CC, CC.VB, CC.DT, PP.RB, CC.JJ, IN.WDT, RB.CC, CC.CD, TO.WDT, CC.VBD, DT.NNS, NNS.CC, RBR.RB and similarly, they share 13 of of the 20 highest rated POS tri-grams: DT.NN.CC, NN.IN.WDT, CC.VB.DT, IN.WDT.PP, NNS.RB.CC,TO.PP.IN, NN.RB.CC, NN.CC.VB, NNS.CC.RB, DT.NNS.CC, RB.DT.NNS, NNS.CC.VB, CC.DT.NNS. Moreover, for word uni-grams and POS bi-grams, the first three features are even in the same order.

The high number of shared items for all three feature types means that case 1 of Distinctiveness prevails, meaning that values for shared features occur with a very different frequency for each author. This is consistent with the clustering results shown in table 6, where for all iterations of cross-validation, the result corresponds to the ideal clustering result.

Figure 5 and figure 6 show the MDS plot and dendrogram corresponding to the shared word uni-gram features of the first iteration (*J-c*). In both one can observe the formation of distinct clusters for both authors with no misclassified document.

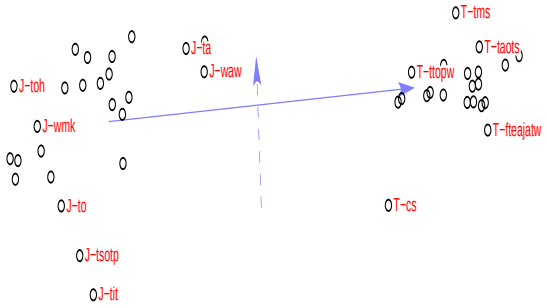


Figure 5: MDS plot on James and Twain highest shared word uni-grams corresponding to iteration 1.

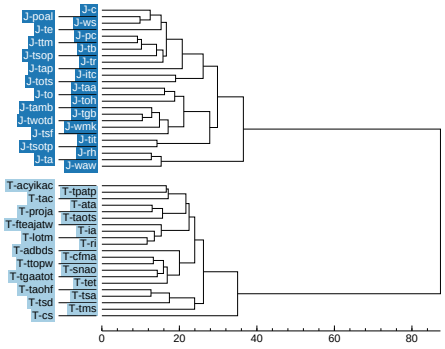


Figure 6: Dendrogram on James and Twain highest shared word uni-grams corresponding to iteration 1

4.3 Discussion

The comparison in both author sets yielded highly distinctive features that were well able to separate the two authors in each study. The second experiment opposing Henry James and Mark Twain was even more successful than the one aiming to find distinctive features of Charles Dickens and Wilkie Collins. There might be various reasons for this difference in accuracy. In terms of stylistics, James and Twain are likely to vary more in style (and topic) than Dickens and Collins. They also wrote for different audiences, where James' books probably appealed mostly to the upper classes, Twain enjoyed general popularity. They also did not approve of each other 's work, which is less likely in Dickens and Collins case, since they collaborated on at least one work. Additionally, one might consider this supported by the fact that while Dickens and Collins were best separated on word uni-grams, with syntactic structures being less well in separation, which might be interpreted as a more intrinsic similarity of building text than James and Twain, for which any of the three feature types returned very high separation results.

However, one has to consider possible outside influences onto the two data sets, such as the prosaic factor of data set size and the time difference. Representativeness and Distinctiveness normalises for the number of comparisons and differences in set size between the compared authors. Yet, there will be a difference between small and bigger sets of an author with respect to a particular feature. If the set is small, it is more probable that values over different documents of an author *agree* for a particular feature, whereas the larger a set becomes, the more different documents have to be *satisfied*. Thus, the slight difference in data set size 54 compared to 46 might have influenced the favourable result for the James/Twain comparison. The other less tangible difference between the sets is time and cultural differences in terms of 'Britishness' vs 'Americanness'.

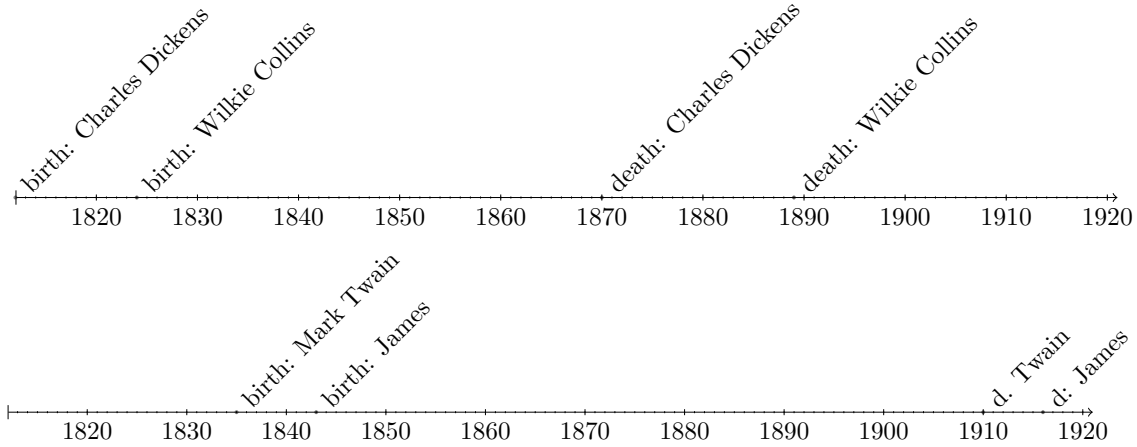

birth: Charles Dickens
birth: Wilkie Collins
death: Charles Dickens
death: Wilkie Collins
1820
1830
1840
1850
1860
1870
1880
1890
1900
1910
1920
birth: Mark Twain
birth: James
d. Twain
d: James
1820
1830
1840
1850
1860
1870
1880
1890
1900
1910
1920

Table 6: Results of Adjusted Rand Index for feature types of word uni-grams, POS bi-grams and POS tri-grams.

Dickens vs. Collins				**James vs. Twain**			
Test Doc (D./C.)	word unig.	POS bigr.	POS trigr.	Test Doc (J./T.)	word unig.	POS bigr.	POS trigr.
D1023	0.93	0.79	0.85	J-c	1	1	1
D1400	0.93	0.66	0.93	J-itc	1	1	1
D19337	0.93	0.79	0.85	J-pc	1	1	1
D40723	0.93	0.79	0.85	J-poal	1	1	1
D564	0.93	0.79	0.93	J-rh	1	1	1
D580	0.93	0.79	0.85	J-ta	1	1	1
D588	0.93	0.79	0.85	J-taa	1	1	1
D650	0.93	0.85	0.93	J-tamb	1	1	1
D675	0.93	0.66	0.79	J-tap	1	1	1
D699	0.93	0.85	0.85	J-tb	1	1	1
D700	0.93	0.79	0.85	J-te	1	1	1
D730	0.93	0.79	0.85	J-tgb	1	1	1
D766	0.93	0.79	0.85	J-tit	1	1	1
D821	0.93	0.85	0.85	J-to	1	1	1
D872	0.93	0.79	0.79	J-toh	1	1	1
D882	0.93	0.85	0.85	J-tots	1	1	1
D883	0.93	0.79	0.85	J-tr	1	1	1
D914	0.93	0.85	0.93	J-tsf	1	1	1
D916	0.93	0.79	0.79	J-tsop	1	1	1
D917	0.93	0.79	0.85	J-tsotp	1	1	1
D963	0.93	0.79	0.85	J-ttm	1	1	1
D967	0.93	0.79	0.85	J-twotd	1	1	1
D968	0.93	0.79	0.85	J-waw	1	1	1
D98	0.93	0.66	0.85	J-wmk	1	1	1
Dal2324	0.93	0.79	0.85	J-ws	1	1	1
Dal28198	0.93	0.85	0.85	T-acyikac	1	1	1
DC1423	0.93	0.79	0.79	T-adbds	1	1	1
C1438	0.85	0.85	0.85	T-ata	1	1	1
C155	0.93	0.79	0.93	T-cfma	1	1	1
C1586	0.93	0.79	0.85	T-cs	1	1	1
C1587	0.93	0.66	0.85	T-fteajatw	1	1	1
C1588	0.93	0.85	0.85	T-ia	1	1	1
C1621	0.93	0.93	0.85	T-lotm	1	1	1
C1622	0.93	0.79	0.85	T-proja	1	1	1
C1623	0.93	0.85	0.93	T-ri	1	1	1
C1624	0.85	0.85	0.85	T-snao	1	1	1
C1626	0.93	0.79	0.72	T-tac	1	1	1
C1627	0.93	0.79	0.85	T-taohf	1	1	1
C1628	0.93	0.79	0.85	T-taots	1	1	1
C1629	0.93	0.85	0.85	T-tet	1	1	1
C1630	0.93	0.79	0.85	T-tgaatot	1	1	1
C170	0.93	0.85	0.85	T-tms	1	1	1
C1895	0.93	0.85	0.85	T-tpatp	1	1	1
C1917	0.93	0.85	0.79	T-tsa	1	1	1
C1975	0.93	0.79	0.79	T-tsd	1	1	1
C28367	0.93	0.85	0.93	T-ttopw	1	1	1
C3606	0.93	0.79	0.79				
C3632	0.93	0.79	0.85				
C3633	0.93	0.79	0.79				
C4605	0.93	0.85	0.85				
C583	0.93	0.66	0.79				
C7891	0.85	0.85	0.79				
C7893	0.93	0.79	0.79				
C7894	0.93	0.79	0.85				

5. Conclusion

References

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

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

Canby, Henry Seidel (1951), *Turn West, Turn East: Mark Twain and Henry James*, Biblo & Tannen Publishers.

Hubert, Lawrence and Phipps Arabie (1985), Comparing partitions, *Journal of classification* **2** (1), pp. 193–218, Springer.

Mahlberg, Michaela (2007), Clusters, key clusters and local textual functions in dickens., *Corpora*.

Mosteller, Frederick and David L. Wallace (2008), *Inference and Disputed Authorship: The Federalist*, The David Hume Series of Philosophy and Cognitive Science Reissues, new ed ed., Center for the Study of Language and Information. http://www.worldcat.org/isbn/1575865521.

Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg, and Therese Leinonen (2011), Gabmap-a web application for dialectology, *Dialectologia: revista electrònica* pp. 65–89.

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

Tabata, Tomoji (2012), Approaching Dickens' Style through Random Forests, *Proceedings of the Digital Humanities 2012*, DH2012.