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 choos- ing 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 along with a method for evaluation of those elements. In order to select an author's consistent features, we propose the measure of Representativeness & Distinctiveness (Prokić et al. 2012) that seeks to identify those elements that are both representative for an author over a given set of his texts, as well distinctive with respect to an opposing author's sample. The method thus bears similarities with both Burrow's Delta and Zeta in favouring consistent terms that are irregular in the opposing author's set. Using the proposed method, we examine different types of features, both lexical and syntactic ones, such as simple word uni-grams, but also Part-of-Speech (POS) bi-grams/tri-grams. For evaluation, we test the separation ability of the selected features by clustering the two authors' documents followed by computing the Adjusted Rand Index (Hubert and Arabie 1985) given the ideal clustering result. 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

Detecting stylistic features in authorship is hampered by the lack of gold standard that would help us to evaluate our findings. In absence thereof, one might resort to methods that are conceived to select consistent features that accomodate and give preference to features that an author uses with a certain regularity over different works and which therefore hint at a clear preference.

Consistency of selection Mosteller and Wallace - strength in numbers - reliability of larger number of stylistic markers

heuristic style of evaluation - looking for desirable characteristics of markers

2. The Data Sets

Investigating two different comparisons of authors, the first set comprising Charles Dickens and contemporary author Wilkie Collins. Interesting, since Dickens' data set contains one collaboration between both authors and two pieces, where Dickens was main author and Dickens was among a group of colloborators. In terms of stylistic properties it will be illuminating to see how these behave with respect to similarity to the other pieces.

2.1 Dickens vs. Collins

2.2 James vs. Twain

Table 1: Dickens' and Collins' data set as part of the Dickens vs. Collins comparison.

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

No.	Author	Texts
1	Collins	After Dark
2	Collins	Antonina
3	Collins	Armadale
4	Collins	Man and Wife
5	Collins	Little Novels
6	Collins	Jezebel's Daughter
7	Collins	I Say No
8	Collins	Hide and Seek
9	Collins	Basil
10	Collins	A Rogue's Life
11	Collins	The Woman in White
12	Collins	The Two Destinies
13	Collins	The Queen of Hearts
14	Collins	The New Magdalen
15	Collins	The Moonstone
16	Collins	The Legacy of Cain
17	Collins	The Law and the Lady
18	Collins	The Haunted Hotel: A Mystery of Modern Venice
19	Collins	The Fallen Leaves
20	Collins	The Evil Genius
21	Collins	No Name
22	Collins	Poor Miss Finch
23	Collins	Rambles Beyond Railways
24	Collins	The Black Robe
25	Collins	Miss or Mrs.?
26	Collins	My Lady's Money
27	Collins	The Dead Alive

Table 2: Twain' and James' data set as part of the Twain vs. James' comparison.

Table 3: 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 4: 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 originally conceived in the realm of dialectometry, where it has been shown to detect lexical items able to distinguish different dialectical areas (Prokić et al. 2012). This study dealt with dialect differences between different sites within a language area with respect to a choice of lexical items. The degree of difference between two sites is characterised by the aggregate differences of comparisons of all lexical items collected at each site. Thus, in the context of dialectrometry, *Representativeness and Distinctiveness* is a measure to detect characteristic features (lexical items), that differ little within a group of sites and considerably more outside that group. Characteristic features are chosen with respect to one group g of sites $|g|$ within a larger group of interest G, where $|G|$ includes the sites s both within and outside g.

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. Thus, 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 input values (eq. 3). The usual input are relative frequencies of the original term frequency weighting, which provide a better picture between the ratio of term frequency and document size. 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 standardized by using all distance values calculated for feature f to yield the degree of representativeness and distinctiveness for term dt in D with respect to DS as defined in eq. 4.

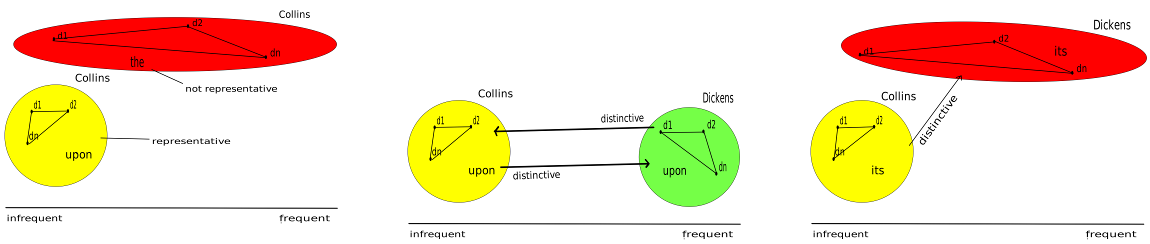
$$dt = \frac{\overline{d_f^{D'}} - \overline{d_f}}{sd(d_f)} - \frac{\overline{d_f^D} - \overline{d_f}}{sd(d_f)} \tag{4}$$

Comparing Authors on the basis of Representative Features In order to evaluate how well the chosen terms do separate the two authors, we motivate the choice of only selecting representative features from both author profiles. The issue in connection with using all discriminatory terms lies at the calculation of the *Distinctiveness* measure. If we calculate representative and distinctive features for an author, we can be sure that the values for those terms are consistently similar for that author, while being different for the outside set. There are consequently two different scenarios with respect to a term *being different* in the other author's set.

1. The term t_i is consistent in set D with a high frequency. The same term t_i is consistent in set the opposing author's set nD with a low frequency. Thus, the term is representative and distinctive for both sets, even though we did not consider the *Representativeness* for set nD. Obviously, the converse could also be true: a consistently low frequency for set D and a consistently high frequency for the set nD. 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.

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

Representativeness — Distinctiveness Case 1 — Distinctiveness Case 2



4. Evaluation through Clustering

Given a list of discriminatory terms for two different author sets, we would like to ascertain to what extent the collection of terms is able to highlight differences between the sets and identify distinct clusters grouping the documents of different authors. As has been shown before, the terms used for discrimination ability should be selected according to separation ability for both author sets. Ideally, frequencies with respect to all terms should be consistent and fairly complementary between two author sets, e.g. Dickens uses *upon* consistently and frequently and Collins uses the term consistently and infrequently. In order to test discrimination ability of a discriminatory term list for two authors, we build a dissimilarity matrix comparing all documents in the complete training set.

4.0.1 Adjusted Rand Index for Evaluation of Clustering

In addition to visual clustering that gives more of an intuition of separation between two sets, a clustering result can be evaluated by comparing two different partitions of a finite set of objects, namely the clustering obtained and the ideal clustering. For this purpose, we can employ the *adjusted Rand Index* (**?**), which is the corrected-for-chance version of the *Rand Index*. Given a set S of n elements, and two clusterings of these points, U and V, defined as $U = \{U_1, U_2, \ldots, U_r\}$ and

$V = \{V_1, V_2, \dots, V_s\}$ with a_i and b_i as the number of objects in cluster U_i and V_i respectively. The overlap between U and V can be summarized in a contingency table 5. where each entry n_{ij} denotes the number of objects in common between U_i *and* V_j : $n_{ij} = |U_i \cap V_j|$.

$$[n_{ij}] = \begin{array}{c} U\ V \\ U_1 \\ U_2 \\ \vdots \\ U_r \\ Sums \end{array} \begin{array}{c} \begin{array}{ccccc} V_1 & V_2 & \dots & V_s & Sums \end{array} \\ \begin{pmatrix} n_{11} & n_{12} & \dots & n_{1s} & a_1 \\ n_{21} & n_{22} & \dots & n_{2s} & a_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ n_{r1} & n_{r2} & \dots & n_{rs} & a_r \\ b_1 & b_2 & \dots & b_s & \end{pmatrix} \end{array} \tag{5}$$

The adjusted form of the *Rand Index* is defined in eq. 6 and more specifically given the contingency table 5 in eq. **??**, where n_{ij}, a_i, b_j are values from the contingency table.

$$AdjustedIndex = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex} \tag{6}$$

The index is bounded between [-1,1], with 0 being the expected value and 1 the highest positive correlation between two different clusterings. For illustration of using the two methods presented above, we consider an example of pairwise comparison of documents of a dataset of $Dickens \cup Collins$, with 55 documents belonging to Dickens and 31 to Collins. This yields a 86 x 86 dissimilarity matrix containing all pairwise comparisons of documents in the set. Figure **??** depicts an example dendrogram showing clustering based on a dissimilarity matrix with distances computed using the *complete link* measure. The *adjusted Rand Index* corresponding to the clustering in figure **??** is 0.82, so very close to the ideal separation, which is also confirmed, when we consider the small number of misclassifications (3 for Dickens and 1 for Collins).

5. Experiments

6. 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.

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

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.