A Diachronic Corpus for Literary Style Analysis

Carmen Klaussner, Carl Vogel
Computational Linguistics Group, Trinity Centre for Computing and Language
School of Computer Science and Statistics
Trinity College Dublin
klaussnc@tcd.ie, vogel@tcd.ie

Abstract

This research presents a resource for diachronic style analysis in particular the analysis of literary authors over time. Temporal style analysis has received comparatively little attention over the past years in spite of the possibility of an author's style frequently changing over time, a property that is not only interesting in its own right, but which also has implications for synchronic analyses of style. The corpus contains 22 American literary authors from the mid 19th to early 20th century that wrote largely in parallel. After describing the resource, we show how the corpus can be used to detect changing features in literary style.

Keywords: Diachronic literary corpus, style analysis, linear regression

1. Introduction

The analysis of the authorial *fingerprint* through computational means, referred to as stylometry, is to be distinguished from manual analyses relying on the verdict of literary scholars particularly acquainted with the author(s) in question. A matter complicating this type of analysis is the fact that an author's style only takes shape through the comparison with other contemporaneous authors, the exact selection of which determines how close one comes to the "actual" fingerprint. For instance, if Mark Twain uses the word *'but'* more frequently than other conjunctions then this is only *interesting* or useful in telling his style apart from others if comparable American authors of his time use the same feature at different rates.

Additionally, one may consider this type of analysis to hold an intrinsic flaw: despite the fact that most authors compose their published writings over a period of 20-40 years, this diachronic property is not widely taken into account. This neglect poses an issue for synchronic style analyses in at least one way. Unless style is found to be invariant for an author and does not change with age and experience, temporality can be a confounding factor in stylometry and authorship attribution (Daelemans, 2013). In the best case, the (synchronic) stylistic analysis might merely select those elements *stable* over time, as this renders them also stable and consistent over the author's corpus, heedlessly discarding those elements that show temporal variation and incidentally possibly style *development*. In the worst case, all features considered are affected by time and the most stable features over the corpus are too variable to discriminate well.

Analysis of diachronic elements of style requires accurately time-stamped data, i.e. either reflecting the time of composition or publication year. This paper describes the development of a parallel literary corpus that allows for comparison among authors as these are both temporally annotated and partially aligned. To the best of our knowledge, this is the first open-source derived literary corpus containing time-stamped texts, which have been collected with the first publication date in mind. Hence, this paper is aimed at presenting this corpus and exemplifying how it can be used for diachronic literary style analysis. More specifically, section 2. considers other diachronic corpora and analyses that have been conducted based thereon. Section 3. describes our corpus specifically and section 4. shows example diachronic style analyses based on this corpus. Section 5. discusses the results and section 6. concludes this work.

2. Previous Research

One of the earlier studies of changes in an author's writing style was the study of the poet William Butler Yeats (Forsyth, 1999). Although using dated texts as a means to develop stable methods for chronological prediction is presented as a main motivation for the study, the question of change in Yeats' style is also mentioned given that scholars do not seem to agree on what change his style is supposed to have undergone. The analysis is based on distinctive marker substrings that are extracted from 142 poems using a modified version of *Monte-Carlo Feature Finding* (a quasi-random search algorithm), which are then ranked according to distinctiveness as measured by χ^2 in separating the categories *Young Yeats* and *Old Yeats*. Poems were divided into these categories based on being written either before or after 1915. Forsyth (1999) reports identifying clear markers of 'young' and 'old' Yeats based on 20 substring markers: for nine out of ten test poems their count is higher in the appropriate age category. In order to be able to assign dates to texts 'a youthful Yeatsian index' is defined as: $YYIX = (YY - OY)/(YY + OY)$, where *YY* refers to the number of younger Yeats markers and *OY* to the number of older Yeats markers found (Forsyth, 1999, p.474). A correlation of $YYIX$ and composition year yields an r of -0.84. When examining two poems that had been revised by Yeats some 30 years later, it is observable that the number of *YY* markers decreased in the revised version, while the number of *OY* markers increased.

Another temporal study was focused on the late 19th century American author Henry James (Hoover, 2007), who is deemed to have changed his style over his creative lifespan (Beach, 1918). Considering the most frequent word unigrams and a variety of different methods, such as Cluster Analysis, Burrows' *Delta*, Principal Component Anal-

Author	Timeline	Gender	Works	Size(MB)
Alice Brown	1884–1922	F	12	5.7
Amanda Minnie Douglas	1866–1914	F	51	24.5
Constance Fenimore Woolson	1873–1895	F	12	6.7
Edith Wharton	1897–1920	F	10	3.5
Elizabeth Stuart Phelps Ward	1866–1907	F	21	5.8
Gertrude Atherton	1888–1923	F	19	9.1
Harriet Beecher Stowe	1852–1886	F	18	11.2
Louisa May Alcott	1854–1893	F	16	5.6
Marion Harland	1854–1914	F	15	9.0
Susan Warner	1850–1884	F	29	18.6
Charles Dudley Warner	1872–1899	M	14	6.1
Edgar Saltus	1884–1919	M	17	3.6
Francis Marion Crawford	1882–1908	M	41	23.3
Harold McGrath	1903–1922	M	15	5.3
Henry James	1877–1917	M	32	17.3
Horatio Alger jr	1866–1906	M	37	10.3
Mark Twain	1869–1916	M	23	11
Robert W. Chambers	1894–1922	M	38	20
Timothy Shay Arthur	1847–1890	M	30	10.7
Upton Sinclair	1898–1922	M	17	8.6
William Dean Howells	1867–1916	M	38	16.7
William Taylor Adams	1855–1896	M	49	17.5

Table 1: Corpus of literary authors, indicating timeline, gender, number of works and their size in megabytes.

ysis and Distinctiveness Ratio, Hoover investigates natural partitions of James' style into three different temporal divisions of early (1877–1881), intermediate (1886–1890) and late style (1897–1917).[1] These three divisions have also been identified by literary scholars (Beach, 1918). Furthermore, Hoover notes the existence of transition periods in between which, for instance, the first novels of the late period being somewhat different from the rest of them. Analysis of the 100 words with the largest Distinctiveness Ratio that are either increasing or decreasing over time show that James appears to have increased in his use of *-ly* adverbs and also in his use of more abstract diction, preferring more abstract terms over concrete ones.

The work on temporal prediction by Klaussner and Vogel (2015) considered the task of accurately predicting the publication year of a text through the relative frequencies of syntactic word features.[2] They used multiple linear regression models to predict the year a text was published in for three data sets, the first containing Mark Twain and Henry James' texts, the second a mid 19th to early 20th reference corpus and a third one combining all data from the previous two sets. Although the two authors' data had been kept separate considering possibly different levels for them, the models disregarding authorial source tended to be more accurate (Root-mean-square-error (RMSE) of 7.2 vs. 8.0 on unseen data).[3] This indicates that Twain and James used their shared features with similar rates. Klaussner and Vogel (2015) also used a reference corpus to examine background language change, specifically *The Corpus of Historical American English (COHA)* (Davies, 2010).[4] They built an accurate model based on this corpus to approximate the general language change over time (RMSE of 4 on unseen data). However, using the same model to predict change in James and Twain was rather inaccurate for both authors (RMSE: 15.4 (Twain) / 20.3 (James)), suggesting that the two authors were rather different to the general language in terms of the stylistic features examined, Twain being somewhat more similar to it than James. Combining all data without reference to authorial source leads to more accurate results (RMSE: 1.8) and model features and estimates suggest a marked influence of Twain and James on the model through change in predictors and their associated weights. Conceptually, this set can be thought of representing a style (change) of a community, where a large proportion of people has a similar style to Twain and James.

All previously described studies have the same serious interpretative issue, i.e. from just examining one to two authors it is not obvious what elements of style change can be attributed to the individual and which would be shared by the larger community of writers to which he or she belongs. The corpus presented in the next section (section 3.) offers the possibility to draw from a set of 22 different authors, thus offering more interpretative background to what is individual and was is general with respect to literary style

[1] Distinctiveness Ratio: Measure of variability defined by the rate of occurrence of a word in a text divided by its rate of occurrence in another. Principal Component Analysis (PCA) is an unsupervised statistical technique to convert a set of possibly related variables to a new uncorrelated representation or principal components.

[2] Syntactic word features are words marked for their syntactic category within context.

[3] Hereafter, when we report RMSE we take the units to be years and do not repeat the unit. This is to be understood with respect to the caveat that the data is processed using only integer values of years. It is not the case, that temporal prediction for any text can be wrong by "7.2 years" - rather by seven years or eight years. The RMSE is an aggregate.

[4] A free sample version is accessible on: `http://corpus.byu.edu/coha` – last verified February 2018.

change.

3. Diachronic corpus

Table 1 shows the set of literary authors, comprising twenty women and twenty-two men, all of whom composed work between 1847–1923.[5] The corpus was populated in the following way: Henry James was chosen based on analyses in the literature that suggested that his style had undergone notable change over time (Hoover, 2007; Beach, 1918). Mark Twain, somewhat of a rival author, presented an interesting contrast to James (Beach, 1918; Canby, 1951), especially since Mark Twain and William James (Henry's brother) maintained an active friendship throughout their lives, both being interested in the Psychical Research and paranormal phenomena.[6]

The remaining authors were chosen by first assembling a list of male and female American authors of the 19th–20th century using *Wikipedia*[7] and then choosing those who had a few works publicly available and spread out over at least twenty years. Also, for the purpose of estimating stable word distributions, it was decided that works had to be at least 150 kilobytes in length thus discarding authors with multiple shorter works. Thus, there might be a bias towards more prominent writers, as there could be more incentive to make their data publicly available. For instance, this may result in a shift towards only certain words or expressions being used more frequently throughout. Also, there is little to no racial diversity in the data set as all authors were white, and even though individuals, such as Harriet Beecher Stowe's writings describe African Americans' conflicts, most authors probably remained in their sphere and wrote predominantly about the type of society they were exposed to themselves. Therefore, any inferences based on this set of literary authors does not necessarily extend to the population of American literary authors at large.

Apart from the apparent dislike James and Twain harboured for each other, there were also more positive connections and collaborations between authors of this corpus. Mark Twain and Charles Dudley Warner wrote *The Gilded Age* together.[8] Elizabeth Stuart Phelps Ward seems to have been an admirer of Harriet Beecher Stowe and referred to her in 1896's *"Chapters from a Life"* as the 'greatest of American women'. Constance Fenimore Woolson, a grandniece of James Fenimore Cooper, quoted William Dean Howells in one of her works and established a friendship with Henry James. Her 1884 'East Angels' is seen as a response to James' 'Portrait of a Lady' (Kreiger, 2005). Susan Warner's 1850's 'Wide, Wide World' has been described as a Feminist *Huckleberry Finn*.[9]

In terms of temporal alignment, a fair subset of the authors wrote largely in parallel. For instance, Harriet Beecher Stowe, Louisa May Alcott, Marion Harland and Susan Warner all have their first work in this corpus within four years of each other (1850–1854).[10] Elizabeth Stuart Phelps Ward and Amanda Minnie Douglas both began writing about 15 years later in 1866. The remainder of the female authors' first contribution is somewhat spread out: Constance Fenimore Woolson (1873), Alice Brown (1884), Gertrude Atherton (1888) and lastly Edith Wharton (1897). As for the male authors, Charles Dudley Warner, Mark Twain, William Dean Howells and Horatio Alger jr also made their first appearance within a few years of each other (1866–1872). The second big wave of male authors' first publication clusters around the 1880s: Henry James (1877), Francis Marion Crawford (1882), and Edgar Saltus (1884). Timothy Shay Arthur and William Taylor Adams started publishing slightly earlier than the rest, 1847 and 1855, respectively, and both remained active for about 40 years. Thus, these earlier time lines still have considerable overlap with most of the other writers in the corpus. An exception to this are Upton Sinclair and Robert W. Chambers, Harold McGrath and Edith Wharton, who only started their career in the 1890s or beginning of the 20th century. However, most authors in this corpus should be comparable in that they composed work over at least 20 years in parallel.

The set of literary authors was mainly collected from *Project Gutenberg (PG)*[11] and supplemented with works from the *Internet Archive (IA)*.[12] Project Gutenberg is the more desirable source given that the data is hand transcribed rather than scanned automatically. However, in this case acquiring data with a time stamp close to the first publication date was essential and for this reason and especially when the equivalent Gutenberg version did not have a time stamp, the Internet Archive version was chosen instead if available. The Internet Archive contains scanned version of books using Optical Character Recognition (OCR), and the quality of the processing varied considerably across books and sponsors. In this a trade-off had to be found, balancing accurate time stamp and quality of processing. Occasionally, when content was very noisy due to OCR errors, files were not included at all. In all cases, the date of a file was decided by taking the first available date, e.g. first copyright or publication date, unless a preface clearly stated that the work had been subjected to explicit revisions. The issue with dating in this case is that both dating a work too early or too late would distort the results.

All data was prepared for processing by manually removing parts that were written at a different time from the main work or introductions or comments not by the author, such as notes or introductions by editors. Additionally, table of contents were also removed, as these do not usually follow a normal sentence structure. Minimal preprocessing was needed for PG files, but the books sourced from the IA could be rather noisy, and as upon inspection each file ap-

[5] The data set is available at `www.scss.tcd.ie/clg/DCLSA/` – last verified February 2018.

[6] `http://www.apa.org/monitor/2010/04/twain.aspx` – last verified February 2018.

[7] `https://en.wikipedia.org/wiki/Category:19th-century_American_writers`– last verified February 2018.

[8] As Twain is listed as first author, it is assigned to his corpus.

[9] Usually, described this way in the book's synopsis.

[10] When using descriptions, such as *first* or *last* with respect to authors' works, this is generally to be understood with respect to this corpus; there might be cases where an earlier or later work for an author exists, but could not be included in this corpus.

[11] `http://www.gutenberg.org/` – last verified February 2018.

[12] `https://archive.org/` – last verified February 2018.

Items Found		Example Context		Occurrence	
Incorrect	*correct*	*incorrect*	*correct*	*raw*	*%*
’11	’ll	you’11	you’ll	15275	0.1
lv	ly	onlv you	only you	154	0.001
n t	n’t	could n t	couldn’t	99465	0.7
}’ / }-	y	exactl}- / exactl}’	exactly	6417	0.05
3011/ 3 r ou / 3ôu	you	3 r ou go home	you go home	2351	0.02
011	no / on	011 the table/ 011 way	on the table / no way	1474	0.01
U	ll / il / li	wiU / wUl	will / will	15895	0.1
/	I / 1 / ! /,	/ will / /”	I will / !”	10067	0.07
AV	W	AVhat	What	4508	0.03

Table 2: Common OCR errors and their correct possible realizations, their raw counts and % of processed IA tokens.

peared to have different types of OCR errors, it was deemed best to correct each file manually to correct scanning errors and remove unwanted formatting sequences. One of the issues with automatically correcting these errors was that even within one file, a misread character could refer to multiple different correct character realizations and only manual examination of the context could accurately determine the correct realization.[13] Errors that have only one possible correct version could be corrected using regular expressions, but manual correction was necessarily in cases where there was more than one possible correct version, e.g. the error ‘011’ could correspond to both ‘no’ and ‘on’ even within the same file. Table 2 shows some of the most common OCR errors and their possible correct realizations as well as occurrence of these and their rates as percentages of the raw corrected tokens in IA texts. All whitespace-separated items in the raw texts add up to 14140296 tokens, which reduces to 13614013 tokens in the manually processed version (a reduction of 4%). We estimated the number of broad differences between the two versions by considering the lines changed compared to all lines in the processed version, i.e. 137594/2146720=0.064 (6.4%).[14] It is important to note that there could be multiple changes per line and simple deletion of superfluous headings or page numbers would not be as time-intensive as manual correction of OCR errors. All processed works add up to 554 files in total, 400 (176.9 MB) from Project Gutenberg and 154 (73.7 MB) from the Internet Archive. When reducing the set to unique *author-publication year* combinations, 409 cases are left.

4. Diachronic Style Analyis

In this section, we present an example of the type of analysis that can be done based on this corpus. Specifically, we examine stylistic change in Henry James and Mark Twain and consider to what extent salient features change in the other literary authors. Section 4.1. briefly introduces the regression technique used to discover linearly changing aspects of style. Section 4.2. then reports on the results.

[13] ‘Character’ here refers to alphanumeric letter.

[14] First, all files were compared pairwise using *diff* in linux, followed by counting changed lines in the resulting output and comparing this to the overall line count in the processed data.

Feature	Model weight
beyond.IN	23515.5
broad.JJ	−37652.9
case.NN	13258.1
feet.NNS	−2693.2
joy.NN	25044.9
other.JJ	7966.1
real.JJ	13550.6
things.NNS	13562.0
usual.JJ	−19171.8
ways.NNS	1468.7
word.NN	7535.7
wore.VBD	−10892.7
since.RB	30815.1
sort.NN	−4496.1

Table 3: Syntactic word features included in the best out of four James-Twain models.

4.1. Methods

For the stylistic feature experiments, we consider ‘syntactic word’ sequences, meaning words that have been marked for their syntactic class, and thus each word is augmented with its respective part-of-speech tag.[15] In cases where an author had more than one work per year, the respective feature token count is collapsed to form a single entry for that year. The following experimental paradigm was first introduced by Klaussner and Vogel (2015) and then further developed by Klaussner and Vogel (2017) to its current state. Thereby, a set of features is selected based on its accuracy in predicting the publication year of a text. Thus, the prediction of a variable y using explanatory models is based on a function over a set of distinct variables: $\{x_1, x_2, \dots x_{p-1}, x_p\} = X$ with $y \notin X$, at the same time point $t : \{t \in 1, ...n\}$ and some error term: $y_t = f(x_{1t} \dots x_{2t}, \dots x_{p-1t} \dots x_{pt}, error)$. The general model for this is shown in eq. 1, predicting variable y, where $\hat{y}_t$ refers to the estimate of that variable at a particular time instance $t : \{t \in 1, ...n\}$, β_0 refers to the intercept

[15] To extract part-of-speech features needed for syntactic word features, the TreeTagger POS tagger (Michalke, 2014; Schmid, 1994) was used.

and β_p to the p-th coefficient of the p-th predictor x_{pt}.

$$\hat{y}_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots \beta_p x_{pt}. \quad (1)$$

In the present case, the 'year of publication' is always set as the response variable, e.g. a model based on syntactic word unigrams (relative frequencies) for the year 1880 could be defined in the following way: $\hat{y}_{1880} = \beta_0 + \beta_1(I.pp_{1880}) + \beta_2(he.pp_{1880}) + \beta_3(a.det_{1880})$.

For this work, *'shrinkage'* models and specifically 'lasso' and 'ridge' as part of the 'elastic net' regression were used (Zou and Hastie, 2005).[16] These models offer an extension to the regular 'ordinary least squares' (OLS) models by additionally penalizing the magnitude of the model coefficients thus aiming to keep the model from overfitting to the data. The elastic net penalizes both the L_1 and L_2 norms.[17]

For evaluation, we used the *Root-mean-square-error* (RMSE): it is defined as the square root of the variance of the residuals between outcome and predicted value and provides the standard deviation around the predicted value, as shown in eq. 2. In the present case, RMSE units would correspond to deviations in years, e.g. a RMSE of 2 translates to an error of 2 years around the actual value.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}} \quad (2)$$

4.2. Experiments

For the experiments, the Twain and James data (40 cases) was separated into training and test sets by using a 75/25 split on the response variable 'year of publication'.[18] Then, the features appearing in all training instances over both their data points were extracted.[19] Using the above regression models in five-fold cross-validation, the best model within 1 standard error (SE) of the model with the lowest error, as defined by the mean-square-error MSE was selected. We ran this configuration four times, constructing different training and test splits each time in order to identify salient features over different divisions. Mean training and test set accuracy over all four iteration are 9.7 and 9.9 RMSE respectively. The best performing model in terms of accuracy has 14 features, shown in table 3, and achieves a RMSE of 9.8 on the training and 5.6 on the test set. Prevalent features over all four iterations are: *"broad.JJ"*, *"case.NN"*, *"other.JJ"*, *"things.NNS"*, *"usual.JJ"*, *"word.NN"* and *"wore.VBD"*.

Figure 1 shows two salient features for James and Twain: the adjectives *broad* and *usual*, where both decrease in usage over time. However, without having examined other authors for the same features, it is not clear whether James' and Twain's common trend is remarkable and indicates agreement among them or is fairly common for American writers at this time. Thus, one of the questions arising from this analysis is whether other contemporaneous authors show a similar trend for these features.

For this purpose, we use our predictive James-Twain model to predict the publication year for each author in the corpus separately. If James and Twain's sharing common trends for these syntactic word features is truly unique, publication dates of other authors' works in the set should not be predicted accurately on average. Table 4 shows the results of this prediction task for each author separately in the second column and the results for only those works within the same year range as the training data in the third column. The final column shows the drop in RMSE with respect to the number of test items left out.[20] Both the works of Twain and James are predicted with a similar accuracy, i.e. a RMSE of 4.4 and 4.5 respectively. Considering now prediction accuracy for the remaining authors in the set using the James-Twain model (first column in table 4) shows that most authors' prediction accuracy is far below that of Twain and James. This suggests that their trend for these features may not have followed the same pattern over time. However, some of the authors composed work some time before or after James and Twain and extrapolation may have caused a drop in prediction accuracy. The third column therefore shows what happens, when works outside of Twain and James' combined timeline (1869–1917) are left out for each author. While some scores stay exactly the same, e.g.

Author	RMSE	RMSE (−ext)	−(RMSE/item)
Twain	4.4	4.4	NaN
James	4.5	4.1	−0.4
Arthur	28.8	17.2	−1.0
S. Warner	24.8	19.0	−0.7
Stowe	26.4	16.0	−1.5
Alcott	**11.9**	**5.9**	**−1.5**
Harland	16.7	14.8	−0.9
Adams	13.4	12.0	−0.1
Douglas	10.3	10.0	−0.1
Ward	17.1	14.4	−0.9
Alger	15.2	15.5	+0.1
Howells	11.8	11.8	+0.0
C.D. Warner	13.8	13.8	NaN
Woolson	13.5	13.1	NaN
Crawford	11.9	11.9	NaN
Brown	21.3	14.7	−2.2
Saltus	14.0	11.6	−1.0
Artherton	28.1	20.8	−1.5
Chambers	**32.1**	**29.4**	**−0.5**
Wharton	25.8	23.7	−0.7
Sinclair	20.5	13.0	−1.5
McGrath	27.5	27.0	−0.2

Table 4: Test RMSE for all authors: showing RMSE for all works (RMSE), only works within model range (RMSE (−ext)) and the drop in RMSE per excluded test item (−(RMSE/item)). Bold printed authors are graphically compared to James and Twain in figure 2.

[16] All regression models were computed using the *glmnet* package in R (Friedman et al., 2010), which in our opinion currently offers the most transparent and flexible implementation.

[17] L1: $\|\beta\|_1$: $\sum_i |\beta_i|$ and L2: $\|\beta\|_2^2$: $\sum_i \beta_i^2$

[18] Using the *caret* package in R (Kuhn et al., 2014).

[19] In further studies, this constraint could be relaxed to 'present in most instances'.

[20] 'NaN' indicates that no removal of test items outside the range was necessary, hence no change in RMSE.

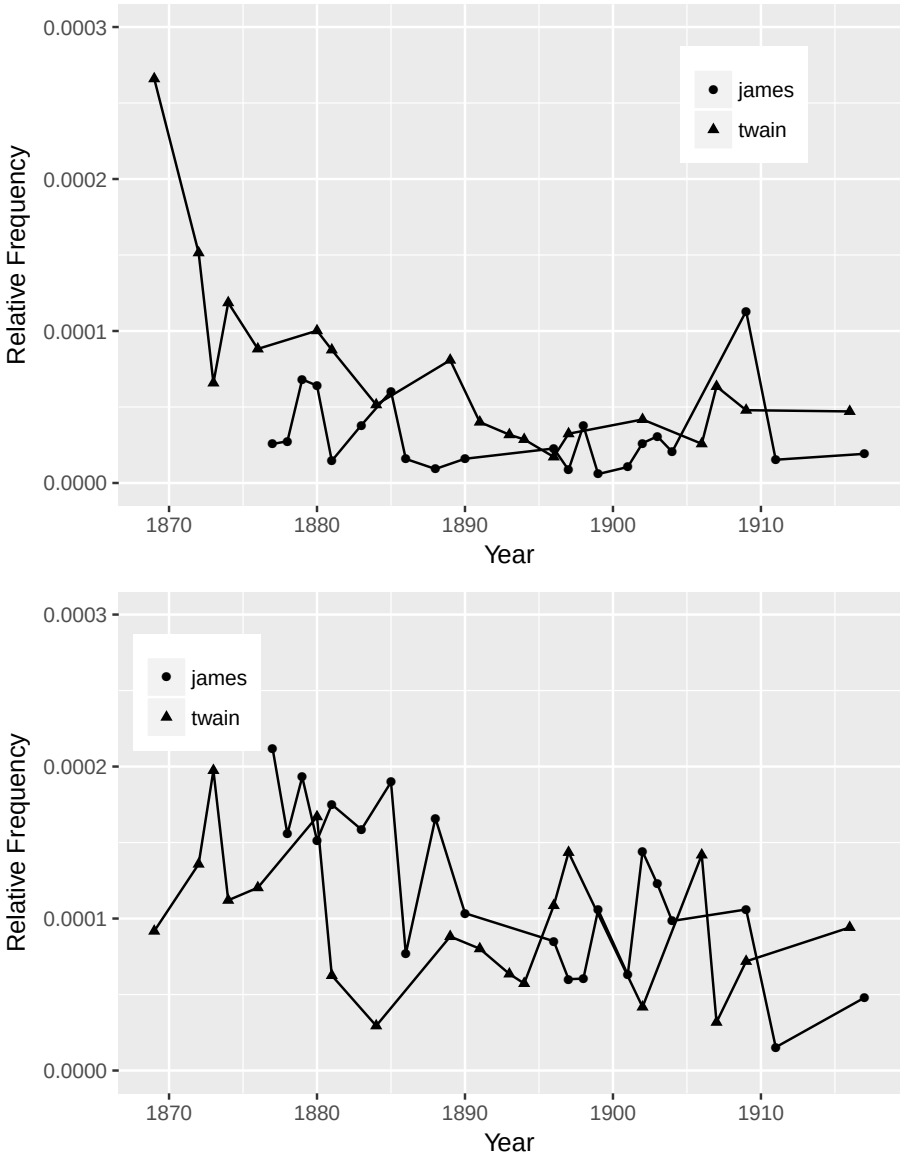

Figure 1: Two salient features for Twain and James. The adjectives 'broad' (top) and 'usual' (bottom).

Adams and Howells, some of them rather improve in accuracy. For instance, leaving out Louisa May Alcott's extra works results in a drop from 11.9 to merely 5.9 RMSE. This actually makes her works fit almost as well with the model as Twain and James' data, even though it was not trained on it. Figure 2 shows pairwise differences between Alcott, Twain and James for the adjective *usual*. Excluding her earlier works before 1865 renders her change in these features a lot closer to the other two authors. In comparison, we examine the author with the highest error, even after removing difficult test pieces, Robert W. Chambers. From the pairwise differences in the bottom plot in figure 2, it can be observed that his trend for the same feature appears to be on a different frequency level than James and Twain, explaining why the model may not be able to accurately date his works based on this high ranking feature among others.

5. Discussion

Section 3. introduced and described a new parallel diachronic literary corpus that can be used to compare among female and male American literary authors from the 19th century. The analysis in the previous section has shown how one can detect salient features based on a two-author set that are discriminatory as to the publication date of texts of these authors. Highly salient model features could be interpreted as being interesting in terms of what these authors have in common when considering stylistic change over time.[21] Yet to what extent these features are changing in this fashion exclusively for the authors considered can only be decided by examining contemporaneous authors that composed works in parallel. Our analysis of works of both Mark Twain and Henry James returned a model with a few highly salient features that when examined visually showed development over time. Trying to use this model to predict other authors' works generally returned much higher error rates than for the two authors on their entire sets. However, the average training and test set error of 9.7 and 9.9 are extremely close to the RMSE scores of

[21] Here, we only considered features that exhibit linear change.

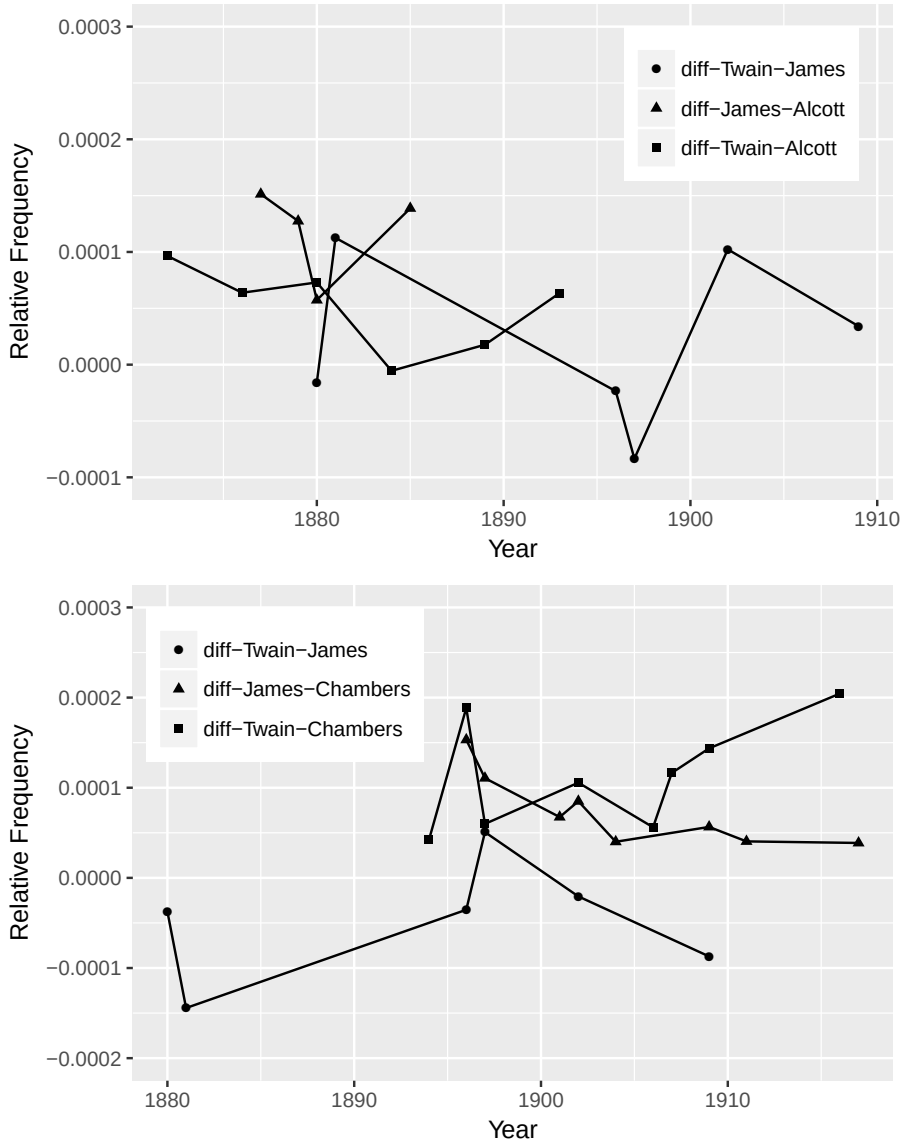

Figure 2: Pairwise differences between Twain and James and Alcott (top) for the adjective 'usual' and Chambers for the noun 'joy' (bottom).

Douglas (10.0), Adams (12.0), Howells (11.8), Crawford (11.9) and Saltus (11.6), especially taking into account that their data would be truly *unseen*. Thus, without also having examined combined models of or with other authors, it is not clear how close James and Twain are in terms of stylistic change regarding the features examined here. This analysis has certainly cast doubt on the *extraordinariness* of shared trends of features, suggesting that parallel analysis of other authors may very well return even a stronger agreement between authors than was witnessed in James and Twain's case.

This analysis has been inclusive with respect to showing that there are common James-Twain features unique in style development to these two authors. Lousia May Alcott gets arguably too close in terms of temporal development to render the discovered features true James and Twain markers. This finding does not necessarily extend to other feature types, for instance they could share unique similarities on stem or syntactic features. What this analysis has shown is that, even for pairwise comparisons, contemporaneous authors need to be examined in parallel in order to give meaning to the individual analyses.

6. Conclusion

In order to analyze style change accurately and determine what features are likely to be more unique in the particular author's case, other contemporary authors have to be examined in parallel. This paper has presented a corpus that can be used for just this purpose, specifically to analyze an author's style with respect to other authors that have composed works during the same time span.

7. Acknowledgements

We would like to thank our anonymous reviewers for their helpful suggestions on how to improve the earlier version of this paper. This research is supported by Science Foundation Ireland (SFI) through the CNGL Programme (Grant 12/CE/I2267 and 13/RC/2106) in the ADAPT Centre (www.adaptcentre.ie).

8. Bibliographical References

Beach, J. W. (1918). *The Method of Henry James*. Yale University Press.

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

Daelemans, W. (2013). Explanation in computational stylometry. In *Computational Linguistics and Intelligent Text Processing*, pages 451–462. Springer.

Davies, M. (2010). The corpus of historical American English: 400 million words, 1810-2009. *http://corpus. byu. edu/coha/.*, 24:2011. (last verified: 21.02.2018).

Forsyth, R. (1999). Stylochronometry with substrings, or: a poet young and old. *Literary and Linguistic Computing*, 14(4):467–478.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Hoover, D. L. (2007). Corpus stylistics, stylometry, and the styles of Henry James. *Style*, 41(2):174–203.

Klaussner, C. and Vogel, C. (2015). Stylochronometry: Timeline prediction in stylometric analysis. In *Research and Development in Intelligent Systems XXXII*, pages 91–106. Springer.

Klaussner, C. and Vogel, C. (2017). Temporal predictive regression models for linguistic style analysis. *Journal of Language Modeling*, 5(3):To appear.

Kreiger, G. (2005). East Angels: Constance Fenimore Woolson's revision of Henry James's the Portrait of a Lady. *Legacy*, 22(1):18–29.

Kuhn, M., Contributions from Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., and the R Core Team, (2014). *caret: Classification and Regression Training*. R package version 6.0-30.

Michalke, M., (2014). *koRpus: An R Package for Text Analysis*. (Version 0.05-4).

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*, volume 12, pages 44–49. Manchester, UK.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320.