Stylochronometry: Timeline Prediction in Stylometric Analysis

Carmen Klaussner and Carl Vogel

Abstract We examine stylochronometry, the question of measuring change in linguistic style over time within an authorial canon and in relation to change in language in general use over a contemporaneous period. We take the works of two prolific authors from the $19^{th}/20^{th}$ century, Henry James and Mark Twain, and identify variables that change for them over time. We present a method of analysis applying regression on linguistic variables in predicting a temporal variable. In order to identify individual authors' effects on the model, we compare the model based on the novelists' works to a model based on a $19^{th}/20^{th}$ century American English reference set. We evaluate using R^2 and *Root mean square error (RMSE)*, that indicates the average error on predicting the year. On the two-author data, we achieve an RMSE of ± 7.2 years on unseen data (baseline: ± 13.2); for the larger reference set, our model obtains an RMSE of ± 4 on unseen data (baseline: ± 17).

1 Introduction

In authorship analysis, it is a natural idealization to treat different works of an author as synchronous events even though this is tantamount to the impossibility that they were all written at the same instant. The assumption is made despite the fact that the works of prolific authors are partially ordered over their lifetimes: some works will have been composed in a non-overlapping sequential manner, while others, largely in parallel over more or less the same duration. Therefore, this takes into account neither the individual changes that an author's style might undergo over time, nor the general underlying language change influencing all contemporaneous writers.

Carmen Klaussner
ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin, Ireland e-mail: klaussnc@tcd.ie

Carl Vogel
Centre for Computing and Language Studies, Trinity College Dublin, Ireland e-mail: vogel@tcd.ie

Ignoring the distinctiveness of an author with respect to other authors, it is relevant to consider the variables that separate each period of composition for an author from other periods for the same author. For example, if we consider an author such as Henry James, who is widely perceived to have changed his style considerably from his early to late works (Hoover, 2007; Beach, 1918), the variables for which he remained consistent might be as interesting to examine as those which may be quantified as having undergone great change. The external factors that may have influenced whether variables are in one category or the other may be of great interest.

Enormous amounts of human ingenuity have been applied over the centuries to the task of temporal classification of text authorship.[1] Methods such as are explored here contribute to semi-automatic methods that draw on text-internal analysis to support stylochronometry. These are generalizations of authorship attribution problems. In the present work, rather than trying to learn features that discriminate two or more authors in synchronic terms, analyzing each one's collection of works against the others' works, we mean to identify elements that are not only prevalent over time, but also provide good indicators of the year a text originated in. In this, one needs to differentiate between individual style change of particular authors as opposed to general language change over time independent of any individual writer. For this purpose, we build regression models based on the works of two prolific authors of the late 19th to early 20th century, Henry James and Mark Twain, as well as models based on a reference corpus corresponding to language use at that time.

In §2, we situate our work with respect to other contributions in the literature. The details of the corpus collection and treatment are outlined in §3. In this section, we also present a methodology for conducting this sort of analysis in general. The data treatment and methods of each individual experiment are outlined in §4, and the results are presented. The outcomes are discussed in §5. Finally, in §6 we conclude.

2 Previous work

Language change is ever present and complicates analysis and comparison of works of different temporal origin. Apart from being of interest in terms of style change over time, this also presents an issue for synchronic analyses of style, as discussed in (Daelemans, 2013): 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.[2] Stamou (2008) reports on various studies in the domain and suggests applying more common methodologies to make comparisons between studies in stylochronometry more feasible.

There have been longitudinal studies on linguistic change with respect to grammatical complexity and idea density, contrasting participants who were to develop

[1] See Coleman (1971) and Frontini et al. (2008) for discussion of attempts in the 15th century to date a text purported to be from the 3rd but shown to be most likely from (circa) the 8th. The former depends on manual methods and the latter, semi-automatic methods.

[2] Early-Wittgenstein may be stylistically as well as conceptually distinct from Late-Wittgenstein.

dementia against those who were not (Kemper et al., 2001), showing that both variables declined over time for both groups although at different rates.

Recent research concentrated on detecting changes in writing styles of two Turkish authors, Cetin Altan and Yasar Kemal, in old and new works (Can and Patton, 2004). The study looked at three different style markers: type and token length and the frequency of the most frequent word unigrams. Employing different methods, such as linear regression, PCA and ANOVA, they found that word types are slightly better discriminators than type and token length.[3] That study is similar to the current one in that it also used regression analysis to evaluate the relationship between the age of a work and particular variables, although token length was used rather than words' relative frequencies as we do here. The authors report a strong relationship between average token length and age of text in Altan's works, although an R^2 value of 0.24 indicates that there are likely to be other factors involved.[4]

Regarding temporal style analysis with respect to an author considered here as well, Hoover (2007) investigates changes in James' style using word unigrams (100–4000 most frequent) with different methods, such as Cluster Analysis, Burrows' *Delta*, Principal Component Analysis and Distinctiveness Ratio.[5] Three different divisions in early (1877–1881), intermediate (1886–1890) and late style (1897–) (that have also been identified by literary scholars (Beach, 1918)) are identified, although there are transitions inbetween with, for instance, the first novels of the late period being somewhat different from the rest of them. The results on the 100 words with the largest Distinctiveness Ratio 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. This work on James' style brought the writer to our attention as an interesting candidate for a temporal analysis of style. In contrast to the previous study, the work we present here focuses on a seamless interpretation of style over time rather than classification into different periods along the timeline of an author's works.

3 Data and Methods

In §3.1, we describe the data sets used and the feature preprocessing applied. We outline a general method for preparing this kind of text data for temporal analysis and introduce time-oriented analysis using explanatory regression models in §3.2.

[3] ANalysis Of VAriance (ANOVA) is a collection of methods developed by R. A. Fisher to analyze differences within and between different groups. 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.

[4] The coefficient of determination R^2 indicates how well a model fits the observed data ranging from 0 to 1 – 0 indicating a poor fit and 1 a perfect one; in the case of evaluating predictions against the outcome (test set) values can also range from -1 to 1; – in the case of negative values, the mean of the data provides a better fit.

[5] 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.

3.1 Corpora

For this study, we consider works of individual authors, Mark Twain and Henry James – both who wrote during the late 19th century to the early 20th century – as well as a reference corpus comprising language of that time. Even though James' and Twain's timelines are not completely synchronous, they largely overlap, which renders them suitable candidates for a combined temporal analysis. In addition, they seem to have been, although both considered to be highly articulate and creative writers, contrasting in temperament and in their art (Canby, 1951, p. xii), yet each conscious of the other (Brooks, 1922; Ayres, 2010). It is interesting to see to what extent perceived differences are apparent in predictive models based on their data.

Table 1 and Table 2 show James' and Twain's main works, 31 and 20 works respectively[6] collected from the *Project Gutenberg*[7] and the *Internet Archive*.[8] Project Gutenberg is the better source in terms of text formatting, but works are not always labelled with publication date, and especially for Henry James, who is known to have revised many works, one has to be sure of the exact version used. Ideally, collected pieces should be close to the original publication date to avoid confounding factors; otherwise, the collected piece might not be the same as the one originally published, and this may introduce irregularities into time-oriented analysis.

The reference corpus is an extract from the *The Corpus of Historical American English (COHA)* (Davies, 2010) which comprises samples of American English from 1810–2009 from different sources, such as fiction and news articles. For the purpose of the current experiments, we consider texts starting from the 1860s to the 1910s in order to cover both authors' creative life span. There are 1000–2500 files for each decade, spread over the individual years and genre. Models built on the basis of this data are likely to be more complete than the authorial data sets, as this collection is more balanced without gaps in the timeline.

In order to extract the features of interest from the texts, we build *R* scripts to lowercase all text before extracting context sensitive word unigrams by using Part-Of-Speech (POS) tagging from the R *koRpus* package (that uses TreeTagger POS tagger) (R Core Team, 2014; Michalke, 2013; Schmid, 1994). Thus, we distinguish between different function/syntactic contexts of one lexical representation: e.g. without taking the context into account, the item LIKE could refer to the verb LIKE or the preposition LIKE. Since we would consider these to be separate entities despite them sharing the same lexical representation, we create separate entries for these, i.e. ⟨LIKE.VB⟩ and ⟨LIKE.IN⟩.[9] Punctuation and sentence endings are also included as features and in relativization (discussed in section 3.2).

---

[6] Here, we only include the main works/novels for reasons of text length and genre homogeneity.

[7] http://www.gutenberg.org/ – last verified August 2015.

[8] https://archive.org/ – last verified August 2015.

[9] The separate entries are created using the POS tags assigned by the tagger to the individual word entity in its context.

Table 1: Henry James' main works. Showing *Title*, the original publication date (*1st Pub.*), version collected (*Version*), *Size* in kilobytes and *Genre*. The dashed lines indicate the boundaries for the compression, i.e. which of the works are combined into one temporal "bucket".

Title	1st Pub.	Version	Size	Genre
The American	1877	1877	721	novel
Watch and Ward	1871	1878	345	novel
Daisy Miller	1879	1879	119	novella
The Europeans	1878	1879	346	novel
Hawthorne	1879	1879	314	biography
Confidence	1879	1880	429	novel
Washington Square	1880	1881	360	novel
Portrait of a Lady	1881	1882	1200	novel
Roderick Hudson	1875	1883	750	novel
The Bostonians	1886	1886	906	novel
Princess Casamassima	1886	1886	1100	novel
The Reverberator	1888	1888	297	novel
The Aspern Papers	1888	1888	202	novella
The Tragic Muse	1890	1890	1100	novel
Picture and Text	1893	1893	182	essays
The Other House	1896	1896	406	novel
What Maisie Knew	1897	1897	540	novel
The Spoils of Poynton	1897	1897	376	novel
Turn of the Screw	1898	1898	223	novella
The Awkward Age	1899	1899	749	novel
Little Tour in France	1884	1900	418	travel writings
The Sacred Fount	1901	1901	407	novel
The Wings of the Dove	1902	1902	1,003.7	novel
The Golden Bowl	1904	1904	1100	novel
Views and Reviews	1908	1908	279	literary criticism
Italian Hours	1909	1909	711	travel essays
The Ambassadors	1903	1909	890	novel
The Outcry	1911	1911	304	novel
The Ivory Tower (unfinished)	1917	1917	488	novel
The Sense of the Past (unfinished)	1917	1917	491	novel
In the Cage	1893	1919	191	novella

3.2 Timeline Compression and Analysis

As can be observed from the data in Table 1 and Table 2, both authors composed works over the span of around forty years each, with overlap for about twenty years. However, in each case works are unevenly distributed with some years giving rise to more than one work. In the present context, where we aim to predict the year on the basis of word features, we combine different works in a year into one.[10] In the following experiments, we sometimes combine all available data for a year or if we investigate different sources (authors) we process these separately and differentiate between them by adding a CLASS attribute indicating the author that is a categorical variable rather than the ordinal YEAR or a continuous lexical variable.[11] Thus, in the context of style analysis, we examine a particular variable *v* over time by considering its relative frequency distribution, e.g. we count the occurrence of that particular

[10] This is without loss of generality to the bag-of-words analysis of texts in which sentence structures are not used subsequent to POS tagging.

[11] Lexical features are continuous here because we use relative frequencies.

Table 2: Collected Mark Twain's works. Showing *Title*, the original publication date (*1st Pub.*), version collected (*Version*), *Size* in kilobytes and *Genre* type. The dashed lines indicate the boundaries for the compression, i.e. which of the works are combined into one temporal "bucket".

Title	1st Pub.	Version	Size	Genre
Innocents Abroad	1869	1869	1100	travel novel
The Gilded Age: A Tale of Today	1873	1873	866	novel
The Adventures of Tom Sawyer	1876	1876	378	novel
A Tramp Abroad	1880	1880	849	travel literature
Roughing It	1880	1880	923	semi-autobiograph.
The Prince and the Pauper	1881	1881	394	novel
Life on the Mississippi	1883	1883	777	memoir
The Adventures of Huckleberry Finn	1884	1885	586	novel
A Connecticut Yankee in King Arthur's Court	1889	1889	628	novel
The American Claimant	1892	1892	354	novel
The Tragedy of Pudd'nhead Wilson	1894	1894	286	novel
Tom Sawyer Abroad	1894	1894	182	novel
Tom Sawyer Detective	1896	1896	116	novel
Personal Recollections of Joan Arc	1896	1896	796	historical novel
Following the Equator: A Journey Around the World	1897	1897	1000	travel novel
Those Extraordinary Twins	1894	1899	1200	short story
A Double Barrelled Detective Story	1902	1902	103	short story
Christian Science	1907	1907	338	essays
Chapters from My Autobiography	1907	1907	593	autobiograph
The Mysterious Stranger (finished by ed.)	1908	1897–1908	192	novel
Is Shakespeare Dead?	1909	1909	121	semi-autobiograph.

word and relativize by the total number of occurrences of all words in that document (or document bin for multiple works in the same temporal span).[12] Building models on the basis of individual authors might lead to less stable models for prediction, since not all years will have given rise to a publication, and the resulting models will need more interpolation than aggregating yearly bins from both authors' works.

This study is motivated by quantitative forecasting analysis that monitors how a particular variable (or variables) changes over time and uses that information to predict how that variable is likely to behave in future (Makridakis et al., 2008). Thus, the (future) value of a particular variable v is predicted by considering a function over a set of other variable values. One differentiates between the use of a *time-series* and *explanatory models*. Time-series analysis considers the prediction of the value the variable v_i takes at a future time point $t+1$ based on a function f over its values (or errors) at previous distinct points of time ($v_i^t, v_i^{t-1} \ldots v_i^{t-n}$), as shown in example 1.

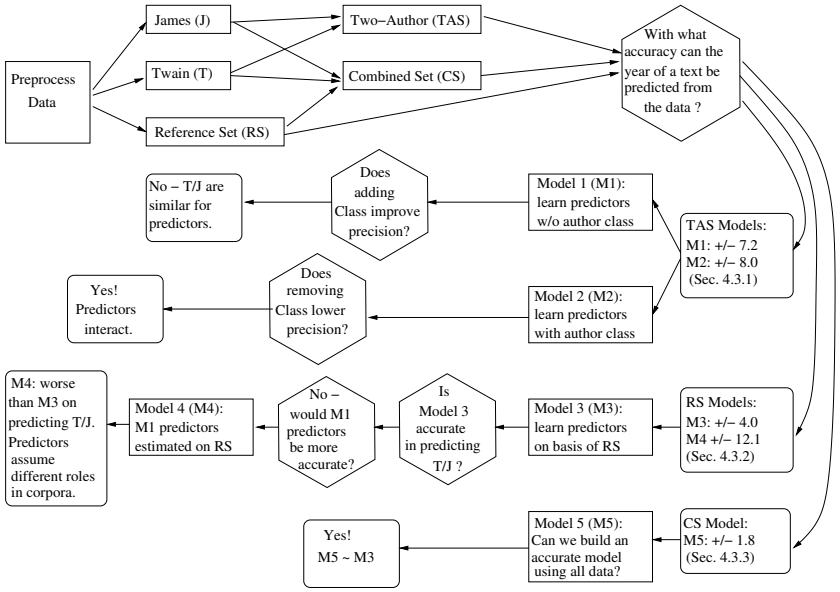
(1) $v_i^{t+1} = \mathrm{f}(v_i^t, v_i^{t-1}, v_i^{t-2} \ldots, error)$

In contrast, *explanatory models* assume that the variable to be predicted has an explanatory relationship with one or more independent variables. Therefore, prediction of a variable v_i is on the basis of a function f over a set of distinct variables: $v_1, v_2 \ldots v_n = V$, with $v_i \notin V$ at the same time point t, as shown in example 2.

(2) $v_i^t = \mathrm{f}(v_1^t \ldots v_n^t, error)$

[12] This applies if it is meaningful to count instances of the variable, as it is for token n-grams: such relativization does not apply, for example, to average word lengths.

Fig. 1: Preparation and sequence of experiments.



Thus, a time-series involves considering prediction on the basis of a variable at distinct time points, explanatory models, which we employ here, consider distinct variables at the same time point; here the latter are taking the shape of multiple regression models predicting the year of publication of a particular text.

4 Experiments

In §4.1, we first present details of the data preparation and the way we constructed the regression models (§4.2). We present our analyses for the data sets in §4.3.

4.1 Data Preparation

In order to build a model to predict the year of a work's publication from the relative frequencies of lexical variables, all data is compressed to an interval level of one year, meaning that counts for features in works of the same year are joined and relativized over the entire token count for that author for that year. In addition, all instances receive a label indicating the YEAR of publication; the two-author data is also marked by a CLASS label. In the case of the two-author model (§4.3.1), empty

years, i.e. those where neither author has published anything, are omitted. This results in thirty-nine cases for all main experiments here. These rows are unique with respect to author and year; there might be cases where both authors have published during the same year, which would result in there being two entries for a particular year; however these are distinct for the CLASS variable. Generally, we only consider features that occur in all YEAR instances in the training corpus to ensure the selection of consistent and regular predictors later on.[13] However, for the two-author experiments, we consider those types that appear in the majority of all instances. Since that data set is much smaller than the reference set, the constant feature selection is more prone to overfit on the training set and would be worse at test set generalization. The reference corpus was preprocessed the same way as the other corpora; however all files belonging to a particular year were joined together, ordering the files arbitrarily, leaving 60 individual year entries spanning from 1860 to 1919 as a basis for calculating feature values.

There are two possible outcomes for selecting features to predict the year of a text in the two-author case. Either the CLASS attribute is among those considered helpful, meaning that there is a difference for authors for the other/some of the variables in the model or it is excluded, indicating that it did not help prediction in combination with the other features selected. Those features are arguably more representative of the language use shared by the authors rather than temporal change in any of the authors considered individually.

For all of the following experiments the data was randomly separated into training and test set to evaluate model generality; the split is 75% and 25% for the training and test set respectively (using the *caret* package in R (Jed Wing et al., 2014)).

4.2 Variable Selection and Model Evaluation

In order to predict the year of a particular work, we consider multiple linear regression models – these however require some pre-selection of features. Even after discarding less constant features, a fair number of possible predictors of about 200-13,000 are left. In order to rank variables according to predictive power with respect to the response variable, we use the *filterVarImp* function in *caret*; this evaluates the relationship between each predictor and the response by fitting a linear model and calculating the absolute value of the t-value for the slope of the predictor.[14] This is

[13] This is not to argue that complementary categories (e.g. relativized counts of features that are not shared between both authors over the entire duration or features that are never shared by the authors over the duration, etc.) are uninteresting. However, for this work we are addressing change in language shared by the two authors and relative to change in background language of their time, thinking that this provides an interesting perspective on their distinctiveness from each other and everyone else.

[14] The t-value measures the size of the difference between an observed sample statistic and its hypothesized population parameter relative to the variation in the sample data. The further the t-value falls on either side of the t-distribution, the greater the evidence against the null hypothesis that there is no significant difference between hypothesized and observed value.

evaluating whether there is a systematic relationship between predictor and response rather than only chance variation. A higher absolute t-value would signal a higher probability of there being a non-random relationship between the two variables.

For the final selection of model predictors, we use *backward* variable selection, whereby the first step tests the full model and then iteratively removes the variable that decreases the error most until further removal results in an error increase.[15] Backward selection might have an advantage over forward selection, which although arguably computationally more efficient, cannot assess the importance of variables in the context of other variables not included yet (Guyon and Elisseeff, 2003). Moreover, some of our exploratory experiments showed that forward selection was more prone to overfitting on the training data.

Model fit is assessed using the adjusted version of the coefficient of determination R^2 (henceforth denoted as: $\bar{R}^2$), which takes into account the number of explanatory variables and thus does not automatically increase when an additional predictor is added; it only increases if the model is improved more than would be expected by chance. R^2 should be evaluated in connection to an F-test assessing the reliability of the result. The F-test evaluates the null-hypothesis that all coefficients in the model are equal to zero versus the alternative that at least one is not – if significant it signals that R^2 is reliable.[16] We also consider the root mean squared error (RMSE), which is the square root of the variance of the residuals between outcome and predicted value.[17] The baseline model for all training/test set divisions is reported on as well; this equates fitting a model where all regression coefficients are equal to zero: this reduces the model to an intercept through the data tested (i.e. the arithmetic mean).[18]

In the following, we only report on models that fulfil the model assumptions measured by the *gvlma* package in R (Pena and Slate, 2014): kurtosis, skewness, nonlinear link function (for testing linearity), heteroscedasticity and global statistics. Thus, any models reported on here will have been found acceptable by this test, and we dispense with reporting acceptability for each individual case.

4.3 Results

Here we present our predictive models; the ones based on only James and Twain are discussed in §4.3.1. Further in §4.3.2, we evaluate two models on the basis of the

[15] In this case, the Akaike information criterion (AIC) is used to evaluate the model: $AIC = -2 * logL + 2k$, where L is the likelihood and k the number of estimated parameters in the model. Thus, AIC rewards goodness-of-fit, but penalizes the number of parameters in the model.

[16] All models reported on here had reliable $\bar{R}^2$ values at a level of a p-value < 0.0001 associated to them, so we dispense with reporting on this in each individual case.

[17] $RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y)^2}{n}}$.

[18] This might not be an entirely realistic scenario in that most predictors, even randomly selected ones, will bear some kind of relation with the response. However, in the case of the test set, the wrong predictors can also have a worse effect than the null-model, so this might be an acceptable approximation.

reference set and in terms of how well they classify works of the individual authors. Finally, we combine both data sets to investigate the effects on the model (§4.3.3).

4.3.1 Two-Author Models

For the first experiment, we consider the lexical features of the two-author training set corpus, which is 273 terms after only retaining features present in most year instances (28 of 31 instances of both Twain (13) and James (18)); the features are then ranked according to predictive power. The baseline model for the training and test data are an estimate of 1892 (RMSE: $\pm$11.3) and 1893 (RMSE: $\pm$13.2) respectively. Thus, the average error in prediction is 11 and 13 years respectively.[19]

One of the best models (a trade-off between training set and test set accuracy) is shown in (3)–this is the result of using the ten highest rated features. $\bar{R}^2$ is 0.71 (RMSE: $\pm$5.5) on the training set and R^2 on the test set is 0.70 (RMSE: $\pm$7.2). All except one predictor are significant with respect to the response variable. In addition, one can check for multicollinearity, i.e. whether the predictors are likely to be correlated: all of them seem only slightly correlated (all values$<$ 2).[20]

(3) *year* = *intercept* + *required.vbn* + *lay.vbd* + *received.vbd* + *put.vbp* + *fail.vb*

In this model, both authors' data was used in unison without taking the individual author of a year instance into account. This implies that the rate at which each of them was using the predictors is unlikely to have been different–these predictors are thus likely to be good indicators of when a piece of text was published, but not necessarily distinctive with respect to either James or Twain. If we manually add the CLASS attribute to the existing model and re-train it, the results change almost imperceptibly on both training and test set by 0.003-0.015 points around 0.70/0.71 for $R^2/\bar{R}^2$ and a 0.2 rise for the RMSE. Thus adding authorship information seems to neither support nor to add conflicting information to the current model. One might interpret this to mean that there is very little difference between the two authors for these predictors. Inspecting the corresponding VIF confirms this in so far as that CLASS does not seem to be particularly related to any of the other predictors.

In order to inspect a model where the CLASS was important, we retain all those features present in 29 of the instances in the corpus (333) and rank these as done previously. The resulting model based on subjecting the best ten features to backward selection is shown in example 4. This model is distinct from the previous one with respect to all predictors. $\bar{R}^2$ on the training set is 0.72 (RMSE: $\pm$5.2) and R^2 on the test set is 0.63 (RMSE: $\pm$8). If we exclude the CLASS attribute from this model, all evaluation parameters deteriorate on both sets. $\bar{R}^2$ drops to 0.62 (RMSE: $\pm$6.3)

[19] The system reports estimates and predictions as decimals; we dispense with reporting these here, as texts were only ordered according to year rather than exact month, which renders those numbers meaningless. R^2 and RMSE are on the basis of rounded versions of predictions.

[20] This can be tested by using the *variable inflation factor (VIF)* that measures how much the variance of the estimated coefficients in regression is inflated compared to when the predictors are not linearly related; a value of 1 to 4 indicating low correlation and 5 to 10 high correlation.

while the test set's R^2 reduces to 0.49 (RMSE: $\pm$9.5). Thus, the CLASS attribute seemed to somewhat interact with the other predictors in the model.

(4) $year = intercept + class + floor.nn + dressed.vbn + blue.jj + waited.vbd + space.nn + sufficiently.rb$

4.3.2 Reference Set Model

Here we investigate how the YEAR is predicted using the reference set rather than the two authors' data. The model is built as before by first creating a random split into training and test data and then discarding features not present in all year instances. The remaining 10,504 features are ranked with respect to the response YEAR and the best five are used in backward selection. The baseline model for training and test set are estimates of 1890 (RMSE: $\pm$17.4) and 1889 (RMSE: $\pm$17) respectively.

The resulting model is shown in example 5. The use of the comma seems to be very telling as it is highly significant as predictor. $\bar{R}^2$ on the training set is 0.96 (RMSE: $\pm$3.2), while R^2 on the test set is comparable with 0.94 (RMSE: c. $\pm$4). There does not seem to be an overlap with the previous models in terms of predictors. Although the model assumptions are met, predictors seem to be somewhat related: $\langle$OUTSIDE.IN$\rangle$ seems to be slightly related to $\langle$,.COMMA$\rangle$; when it is dropped from the model, the VIF of $\langle$,.COMMA$\rangle$ decreases by at least 2 points. This could indicate that these form common collocations, however, this would have to be quantified as part of a concordance analysis.

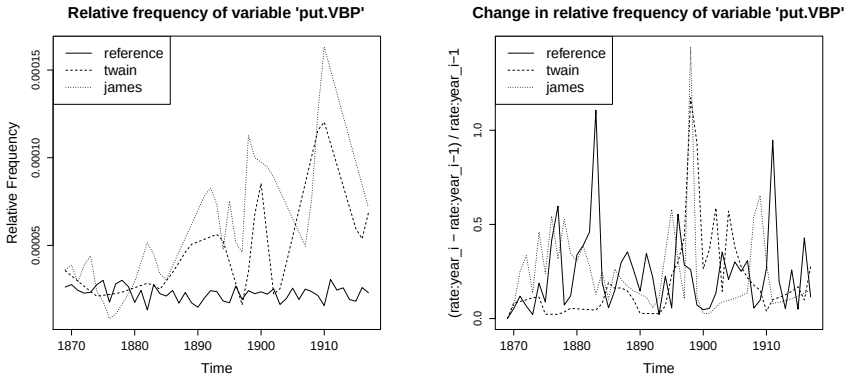
(5) $year = intercept +, .comma + later.rbr + outside.in + planned.vbn$

One question that emerges from this is to what extent the reference model is able to classify James' and Twain's works. Taking each author's year averages separately as test sets (16 for Twain and 23 for James), the reference set model performs quite poorly for both; R^2 is -0.79 (RMSE: $\pm$15.4) / R^2 is -2.1 (RMSE: $\pm$20.3) for Twain and James respectively. The baseline model for James' and Twain's sets are 1889 (RMSE: $\pm$11.5) and 1894 (RMSE: $\pm$11.5) respectively. In this case, the mean through the data provides a better prediction than the reference model.

The predictors that are most reliable for estimating the year for the reference set are not effective for Twain or James. There might be common reliable predictors, but these are not among the ones chosen for the reference set alone, it seems. In this, James' data seems to be harder to classify than Twain; both his scores are considerably worse than Twain's–this might be an indication that James' works differ more from the general style of that time. In order to see whether the reverse is true; reliable predictors for YEAR on the basis of James' and Twain's corpus performing worse on the reference corpus, we use the very first model's predictors to build a model based on the reference corpus data. Thus, the predictors are the same, but the instantiations might be different because of possible deviations in terms of word frequencies. Considering the results of $\bar{R}^2$ of 0.47 (RMSE: $\pm$11.9) on the training set and R^2 on the test set of 0.49 (RMSE: $\pm$12.1) indicate that Twain's and James' predictors are less successful for the reference set data.

Again, taking the two-author's data as test sets returns even worse results than previously: R^2 decreases to $-12/-14$ (RMSE = c. $\pm 42.2/44.7$). Inspecting the model parameters shows that the estimates for the predictors are quite different for the two-author model and the reference corpus; thus these seem to be taking on genuinely different roles in each corpus; this is further depicted in figure 2, where we show one Twain/James predictor over time for each subcorpus separately, showing considerably more variation for Twain and James (partly interpolated).

Fig. 2: Depicting predictor frequency and change of this predictor across all three corpora.



4.3.3 Combining Models

Here, we present a final model built on all the data available, i.e. the reference corpus and the two author data. Thus, all data is aggregated together without reference to the source – James' and Twain's individual year data is added to that of the reference corpus before relativization. After discarding features not in all year instances, 13,245 features remain. As would be expected, adding more data should yield more constant instances than before; thus James and Twain might have constant features that are not present in all of the reference corpus data. The addition of their data contributed to a rise of c. 2,700 more constant features that would not have been constant over the reference corpus on its own. The baseline model for the training set here is the same as for the reference model, as we are only considering the average year over the data sets rather than any features within, the estimates also do not change from the previous ones. Considering the same number of highest ranked features as in the previous reference set model yields the model shown in example 6. This is rather similar to the previous one, except for the features $\langle$WANT.NN$\rangle$ and $\langle$ATTITUDE.NN$\rangle$ rather than the predictor $\langle$PLANNED.VBN$\rangle$. The model's $\bar{R}^2$ on the training set is slightly higher than previously: 0.97 (RMSE: ± 2.8). The test set's R^2 is also higher with 0.988 (correspondingly RMSE: ± 1.8).

Thus, James' and Twain's data might be adding different information in terms of constant features that complement the reference set. The model estimates are somewhat different from previously indicating that the two author data might be creating a shift there as well. Increasing the number of input features causes an improvement on the training set, but slightly less accurate results on the test set.

(6) $year = intercept + , .comma + later.rbr + outside.in + attitude.nn + want.nn$

5 Discussion

The results of these experiments show it possible to accurately predict the year of a publication in the two-author case and in particular in the case where we have a larger (reference) corpus at our disposal. The exact predictor selection is subject to the underlying data set, although the more data is available, the more stable this process seems to become. The results obtained seem to indicate that the model built on the basis of the two-authors (§4.3.1) has to approximate two potentially rather different styles. Using a more balanced corpus in terms of authors and genre seems to create a better approximation to a general style of that time. In order to truly account for the differences between models only built using James and Twain and those built on the larger reference set, one would need to examine the development of those features within in detail in order to see in what way the individual authors deviate from the *general* style. Future work should address those features not attested in all yearly bins in order to investigate differences to constant features examined here as well as individual and general language change, i.e. are some features abandoned over time and does this happen gradually or abruptly. Apart from the word features examined here, one might also consider syntactic shift and in what way prolific authors, such as James and Twain differ from the general style.

6 Conclusion

The stylochronometric analysis reported here supports qualitative assessments of the texts analyzed: despite differences noted between James and Twain, when using their novels to predict year of authorship, their mutual discriminability dissipates. A contribution of this work is to introduce methods of preparation and analysis for the temporal analysis of stylometry. We have shown that it is possible to predict the year of a publication relatively accurately from lexical features whether one is analyzing individual authors or a general reference set of the time. Future work includes the analysis of structural patterns, general and individual ones.

Acknowledgements This research is supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at Trinity College Dublin.

References

1. Ayres, A. *The wit and wisdom of Mark Twain*. Harper Collins, 2010.
2. Beach, J. W. *The Method of Henry James*. Yale University Press, 1918.
3. Brooks, V. W. "The Ordeal of Mark Twain". In: *London: William Heineman* (1922).
4. Can, F. and Patton, J. M. "Change of writing style with time". In: *Computers and the Humanities* 38.1 (2004), pp. 61–82.
5. Canby, H. S. *Turn West, Turn East: Mark Twain and Henry James*. Biblo & Tannen Publishers, 1951.
6. Coleman, C. *The Treatise Lorenzo Valla on the Donation of Constantine: Text and Translation*. First published 1922. New York: Russell & Russell, 1971.
7. Daelemans, W. "Explanation in computational stylometry". In: *Computational Linguistics and Intelligent Text Processing*. Springer, 2013, pp. 451–462.
8. Davies, M. "The Corpus of Historical American English: 400 million words, 1810-2009". In: *http://corpus. byu. edu/coha/*. 24 (2010). (last verified: 24.08.2015), p. 2011.
9. Frontini, F., Lynch, G., and Vogel, C. "Revisiting the Donation of Constantine". In: *2008 Artificial Intelligence and Simulation of Behavior – Symposium: Style in Text*. Ed. by Kibble, R. and Rauchas, S. 2008, pp. 1–9.
10. Guyon, I. and Elisseeff, A. "An introduction to variable and feature selection". In: *The Journal of Machine Learning Research* 3 (2003), pp. 1157–1182.
11. Hoover, D. L. "Corpus Stylistics, Stylometry, and the Styles of Henry James." In: *Style* 41.2 (2007).
12. Jed Wing, M. K. C. from et al. *caret: Classification and Regression Training*. R package version 6.0-30, (last verified: 24.08.2015). 2014. URL: http://CRAN.R-project.org/package=caret.
13. Kemper, S. et al. "Language decline across the life span: findings from the Nun Study." In: *Psychology and aging* 16.2 (2001), p. 227.
14. Makridakis, S., Wheelwright, S. C., and Hyndman, R. J. *Forecasting methods and applications*. John Wiley & Sons, 2008.
15. Michalke, M. *koRpus: An R Package for Text Analysis*. Version 0.04-40, (last verified: 24.08.2015). 2013. URL: http://reaktanz.de/?c=hacking&s=koRpus.
16. Pena, E. A. and Slate, E. H. *gvlma: Global Validation of Linear Models Assumptions*. R package version 1.0.0.2, (last verified: 24.08.2015). 2014. URL: http://CRAN.R-project.org/package=gvlma.
17. R Core Team. *R: A Language and Environment for Statistical Computing*. (last verified: 24.08.2015). R Foundation for Statistical Computing. Vienna, Austria, 2014. URL: http://www.r-project.
18. Schmid, H. "Probabilistic part-of-speech tagging using decision trees". In: *Proceedings of international conference on new methods in language processing*. Vol. 12. Manchester, UK. 1994, pp. 44–49.
19. Stamou, C. "Stylochronometry: stylistic development, sequence of composition, and relative dating". In: *Literary and linguistic computing* 23.2 (2008), pp. 181–199.