

Aspects of linguistic ageing in literary authors across time

Carmen Klaussner, Carl Vogel, and Arnab Bhattacharya
Trinity College Dublin

ABSTRACT

This work offers an investigation into linguistic changes in a corpus of literary authors hypothesized to be attributable to the effects of ageing. In part, the analysis replicates an earlier study into these effects, but adds to it by explicitly analyzing and modelling competing factors, specifically the influence of background language change. Our results suggest that it is likely that this underlying change in language usage is the primary force for the change observed in the linguistic variables that was previously attributed to linguistic ageing. However, our results are tentative insofar as we do not examine non-linear models in general, or other variables influenced by ageing, or non-professional writers who may be more susceptible to these observed shifts in general language than was observable for the literary authors.

Keywords:
linguistic ageing,
diachronic literary
analysis,
language change

INTRODUCTION

1

Language is subject to constant change, both with respect to a particular linguistic variety that affects all its speakers as well as on an individual level for each speaker separately during their lifetime. “Stylometry” is the study of a writer’s stylistic fingerprint based on collected writings over his or her lifetime. Due to the sequential and long term nature of publishing, stylometric studies may be influenced by

temporal language development and its effects might be misconstrued and misinterpreted as a result. More recently this issue has given rise to a temporal variant of stylometric analysis, i.e. “stylochronometry”, that studies changes in style over time, as exemplified, for instance, by the work of Forsyth (1999), Stamou (2007) and Klaussner and Vogel (2018b). However, even though stylochronometric studies consider the temporal dimension, these analyses still conflate individual stylistic changes with those induced by ageing, such as changes in authors’ vocabulary size over time and, most importantly, influences that affect speakers of the same language variety equally, such as general underlying language shifts. Although previous studies have examined sets of linguistic variables with respect to both healthy and pathological ageing (Pennebaker and Stone 2003; Le *et al.* 2011; Kemper *et al.* 2001), to the best of our knowledge there do not yet exist composite studies considering all three aforementioned factors.

The current work extends a research paradigm created by Pennebaker and Stone (2003) who analyzed linguistic ageing both in emotional disclosure studies and in a corpus of literary authors. In this work, we build on previous results by examining a larger literary corpus as well as controlling for background language change. Our objective is to replicate the earlier study on a different literary corpus that is temporally-aligned with a reference corpus for that same time period, thus allowing us to investigate possible influences of general language shifts. We also propose some methods that can be used to attempt to disentangle general effects from those that are individual.

Within this paper, Section 2 discusses previous work in the area, Section 3 presents the literary authors data set, Section 4 and Section 5 discuss methods and experiments respectively and Section 6 and Section 7 analyze and summarize the results.

2

RELATED WORK

Patterns of general, underlying language change have been studied by, for instance Lieberman *et al.* (2007), finding that the question of whether an irregular verb in English will acquire the “-ed” regularization largely depends on its token frequency. Highly entrenched,

irregular verbs such as *have* or *be* are less likely to be regularized. Štajner and Mitkov (2011) investigated diachronic changes in American (AE) and British English (BE) with respect to four different variables: Average Sentence Length (ASL), Automated Readability Index (ARI), Lexical Density (LD) and Lexical Richness (LR) across four different text categories (press, general prose, learned and fiction)^{1,2} Based on two-tailed t-tests, they report a statistically different increase for ARI in BE press/prose (interpreted by the authors as a tendency to render texts more difficult to read in these categories), while the ASL for BE did not change significantly in the period of 1961–1991. Both LR and LD increased across each of the press, prose and fiction categories. In comparison, while AE does not exhibit a significant change in ARI, ASL decreased significantly for the press and learned text categories, which is interpreted as an example of colloquialization. LR and LD in AE only increased in the prose text category. Statistically significant differences between 1961 AE press and 1961 BE press for ARI/LD/LR disappeared by 1991/1992 which is attributed to the growing Americanization that would be particularly tangible in this category.

One of the first statistically-oriented studies into changes in an author's writing style was Forsyth's (1999) study of the poet W. B. Yeats. The study's objectives were to develop stable methods for chronological prediction as well as to examine possible changes in Yeats' style, the exact manifestation of which is disputed among literary scholars. The analysis considered distinctive marker substrings extracted from 142 poems using a modified version of "Monte-Carlo Feature Finding" (a quasi-random search algorithm). Features were then ranked according to distinctiveness measured by χ^2 in separating the categories "Young Yeats" (before 1915) and "Old Yeats" (after 1915). Forsyth (1999) reported 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.

Another literarily-motivated analysis (Hoover 2007) considered the late 19th century American author Henry James, who supposedly

¹ Published: 1961 (BE + AE) and 1991 (BE)/1992 (AE).

² Lexical density is defined as 'number of unique tokens/total number of tokens', whereas lexical richness is defined as 'number of unique lemmas/total number of tokens'.

changed his style over his creative lifespan. Based on literary scholars' findings (e.g. Beach 1918), Hoover investigates natural partitions of James' style into three different temporal divisions of early (1877–1881), intermediate (1886–1890) and late style (1897–1917) using the most frequent word unigrams and a variety of different methods, such as Cluster Analysis, Burrows' Delta, Principal Component Analysis and Distinctiveness Ratio.³ Apart from these divisions, Hoover also notes the existence of gradual transitions in between, with for instance the first novels of the late period being somewhat different from the rest of them.

Le *et al.* (2011) contrasted the writings of three female British novelists for detecting markers of dementia, specifically Iris Murdoch, who died with Alzheimer's disease, Agatha Christie, who was suspected of having it, and P. D. James, who aged healthily. Previous research (Kemper *et al.* 2001; Bird *et al.* 2000; Burke and Shafto 2008 as cited by Le *et al.* 2011) indicated that for instance vocabulary and syntactic complexity declined more rapidly in the presence of dementia, particularly with respect to words of lower frequency and higher specificity as well as passive constructions. Simultaneously, occurrence of lexical repetitions and disfluencies would increase. Analyzing a variety of lexical and syntactic measures, Le *et al.* (2011) could largely confirm their hypotheses with regard to more rapid lexical decline in Murdoch. More than 20 years before any Alzheimer's symptoms became apparent, her vocabulary started to decline, resulting in a significant increase in lexical repetitions of content words. However, her lexical specificity, measured through the proportion of specific indefinite nouns and verbs, remained intact throughout. All but two of Christie's lexical types showed an overall decline. In contrast, the vocabulary, repetition and specificity scores vary only slightly across James' novels. Thus, it is noted that although Murdoch does not share Christie's increase in indefinite nouns, they both show common lexical decline not found in James, validating the hypotheses with respect to lexical markers.

Although the analysis and data preparation was very carefully conducted, as the authors note, the data set is somewhat small with

³Distinctiveness Ratio is a measure of variability defined by the rate of occurrence of a word in a text divided by its rate of occurrence in another.

only 1–2 people for each of the two conditions, leaving it unclear what aspect of the results are reliable, as each of the examined women could potentially be unrepresentative of their group. In addition, general language shift or stylistic change could also have had an influence on the observed change.

While the work by Le *et al.* (2011) considered symptoms of pathological linguistic decline, the study by Pennebaker and Stone (2003) (hereafter also: P&S) focused on aspects of regular and expected linguistic ageing. In particular, they proposed four hypotheses about the effect of ageing on language. Firstly, they suggested that ageing was associated with a drop in negative affect words and a slight increase in positive affect words (hypothesis 1). Further, social words and first-person plural pronouns⁴ were hypothesized to decrease relative to a person's decrease in social networks (hypothesis 2). If ageing was associated with a greater concern with the past relative to the future, linguistic shifts from future to past tense as well as a reduction in references to time altogether could be expected (hypothesis 3). Finally, older people were predicted to use fewer cognitively complex words (cognitive mechanisms and causal, insight, and exclusive words), whereas markers of verbal ability were not expected to show either monotonic increases or decreases (hypothesis 4). P&S investigated how the age of a person affected these linguistic categories, with respect to two very different data sets: one based on self-reports from emotional disclosure studies (the 'Disclosure project'; hereafter also: DP) and the other based on collected works of ten different authors across their individual life spans, hereafter also referred to as the 'Author project' (AP).

The Disclosure project featured 3,280 participants from 45 separate studies, of which 32 were traditional emotional disclosure experiments in which participants were randomly assigned to write about either a traumatic or emotional topic, or a superficial topic in the case of the controls (for details, see Pennebaker and Stone 2003). Although this data is ordered by age of participants, the samples may have originated from the same time period. Both DP and AP were assessed using correlation analysis and in addition the DP was also analyzed through simple linear and quadratic regression.

⁴These are hereafter also referred to as '1PL'.

Table 1: P&S's results showing means over individual age-variable correlations. Significance t-tests are based on means of the within-author (individual variable) correlations with age for the Author project and between-subject with age for the Disclosure project. Significance levels are indicated by: *: $p \leq 0.05$ /**: $p \leq 0.01$ /***: $p \leq 0.001$

LIWC variable	Example	AP	DP	
			Experimentals	Controls
Social and identity				
First-person singular	<i>I, me, my</i>	-0.26*	-0.18*	-0.18*
First-person plural	<i>we, us, our</i>	0.03	-0.01	-0.27*
Time orientation				
Past-tense verbs	<i>was, went, ate</i>	0.08	-0.20*	-0.22*
Present-tense verbs	<i>am, see, goes</i>	0.09	0.05	0.03
Future-tense verbs	<i>will, shall</i>	0.22*	0.19*	0.10*
Cognitive complexity				
Big words (> 6 letters)	<i>pontification</i>	0.10	0.35*	0.36*

Table 1 shows P&S's results for individual age-variable correlations for both data sets (limited to those variables that are also analyzed as part of the current research, as this paper only details a partial replication of the original study). For this, the results for the two DP conditions ('Experimentals'/'Controls') were based on between-subject analyses correlating each of the Linguistic Inquiry and Word Count (LIWC) variables with age. For the Author project, the correlation coefficient is based on mean within-author correlations between each author's age and the LIWC analyses for the works written at that age. As can be observed from the table, for first-person singular pronouns,⁵ present and future tense, and big words, hereafter also referred to as long-letter sequences, all three correlations are in the same direction across the two sets, although only in two cases are all of them also significant (and the direction of first-person singular pronouns is inversely correlated with age while the other two variables directly correlate with age).

⁵These are hereafter also referred to as '1SG'.

Table 2 shows the collection of authors in the AP of the P&S study. Although it is balanced across genders, it contains some idiosyncrasies, such as the fact that most authors originated from Great Britain (England and Scotland), except for writers Louisa May Alcott and Edna St. Vincent Millay of American origin. Genre types include novels, plays and poetry, a fact that could present a confounding factor specifically for the analysis of pronouns that are usually distributed somewhat differently across these text types. The most relevant issue in this context is that authors' works are spread across five centuries (1591–1939) and language use would be expected to somewhat vary between the 16th and 20th centuries. It is to be assumed that this design was deliberate in order to extract very diverse samples – nevertheless, this may render them still less comparable and results could be spurious. In particular, if language has been affected by a continuous shift throughout this time, a significant effect in authors who did not compose language in parallel may still be attributable to general language change rather than ageing.

The final column in Table 2 shows the result of using regression weights for the LIWC variables based on the DP data to create an ageing coefficient for each individual author, which was then correlated with age. Thus, larger correlations signify more similarity to the DP analysis regarding the ageing variables. It is noticeable that five out of six significant correlations, i.e. Joanna Baille, Robert Graves, Edna St. Vincent Millay, William Wordsworth and William Butler Yeats, are based on genre types that could be more prone to irregularities, e.g. poetry and plays. Overall, neither analysis anchored in the DP or AP data is reported to have evaluated the influence of general language change.

Thus, apart from general language shifts, other possible confounding factors for the P&S study could have been introduced by the differences in pronoun distributions across varying text types as well as individual stylistic differences and developments, irrespective of any particular ageing process. In this work, we revisit the question of linguistic ageing for six variables previously analyzed. Specifically, we do not reanalyze P&S's data, but conduct a comparable experiment on a more temporally and genre-homogenous data set. We then compare our findings on the same variables to P&S's earlier results.

Table 2: P&S's 'Characteristics of Authors Chosen for the Author Project' (Pennebaker and Stone 2003, p. 297). F = female; M = male. The ageing coefficient correlations are within-subject simple correlations between each author's age and the ageing coefficient and were based on the regression weights from the Disclosure Project. Significance levels are indicated as follows: †: $p \leq 0.08$ /*: $p \leq 0.05$ /**: $p \leq 0.001$

Author	Nationality	Sex	Life span	Productive years	Genre	Analyzed works (n)	Words per work (M)	Ageing coefficient correlation
<i>Louisa May Alcott</i>	US	F	1832–1888	1854–1886	Novels, stories	19	40,273	-0.05
<i>Jane Austen</i>	England	F	1775–1817	1787–1817	Novels, stories	13	68,120	0.23
<i>Joanna Baillie</i>	Scotland	F	1762–1851	1789–1827	Plays	20	18,921	0.60**
<i>Charles Dickens</i>	England	M	1812–1870	1836–1870	Novels	15	257,777	-0.23
<i>George Eliot</i>	England	F	1819–1880	1859–1876	Novels, stories	10	157,751	0.63*
<i>Robert Graves</i>	England	M	1895–1985	1910–1975	Poetry	100	1,689	0.18†
<i>Edna St. Vincent Millay</i>	US	F	1892–1950	1917–1947	Poetry	21	3,850	0.72**
<i>William Shakespeare</i>	England	M	1564–1616	1591–1613	Plays	37	22,975	0.03
<i>William Wordsworth</i>	England	M	1770–1850	1785–1847	Poetry	64	6,074	0.37**
<i>William Butler Yeats</i>	England	M	1865–1939	1889–1939	Poetry	34	2,217	0.40*

Note: For most novels, stories, and plays, each work was analyzed separately. For poetry, a work was defined by the various poems written within a given year. Exceptions include poems or collections that were known to have been written over several years, which were entered as separate text files.

The data analyzed for this research is divided into two main sets: twenty-two literary authors, comprising ten women and twelve men, and a corresponding reference corpus for the same time period. Table 3 shows the set of literary authors, all of whom published work between 1847–1923.⁶ The corpus was populated in the following way: first

Table 3: Corpus of literary authors, indicating timeline, gender, number of works, size of works in megabytes and their total word count

Author	Timeline	Gender	Works	Size (MB)	Word count
<i>Alice Brown</i>	1884–1922	F	12	5.7	1064566
<i>Amanda Minnie Douglas</i>	1866–1914	F	51	24.5	4500421
<i>Constance Fenimore Woolson</i>	1873–1895	F	12	6.7	1204937
<i>Edith Wharton</i>	1897–1920	F	10	3.5	609351
<i>Elizabeth Stuart Phelps Ward</i>	1866–1907	F	21	5.8	1055611
<i>Gertrude Atherton</i>	1888–1923	F	19	9.1	1628163
<i>Harriet Beecher Stowe</i>	1852–1886	F	18	11.2	2049014
<i>Louisa May Alcott</i>	1854–1893	F	16	5.6	1027950
<i>Marion Harland</i>	1854–1914	F	15	9.0	1572983
<i>Susan Warner</i>	1850–1884	F	29	18.6	3467028
<i>Charles Dudley Warner</i>	1872–1899	M	14	6.1	1088452
<i>Edgar Saltus</i>	1884–1919	M	17	3.6	650825
<i>Francis Marion Crawford</i>	1882–1908	M	41	23.3	4238660
<i>Harold McGrath</i>	1903–1922	M	15	5.3	945365
<i>Henry James</i>	1877–1917	M	32	17.3	3123582
<i>Horatio Alger</i>	1866–1906	M	37	10.3	1840445
<i>Mark Twain</i>	1869–1916	M	23	11	1990085
<i>Robert W. Chambers</i>	1894–1922	M	38	20	3465933
<i>Timothy Shay Arthur</i>	1847–1890	M	30	10.7	1933432
<i>Upton Sinclair</i>	1898–1922	M	17	8.6	1572977
<i>William Dean Howells</i>	1867–1916	M	38	16.7	3063271
<i>William Taylor Adams</i>	1855–1896	M	49	17.5	3208971

⁶The corpus is motivated and described in more detail by Klaussner and Vogel (2018a). The data set is available at <http://www.scss.tcd.ie/clg/DCLSA/> – last verified October 2021.

the prolific authors Mark Twain and Henry James were chosen, which was inspired by several sources that suggested they may be interesting to contrast (Beach 1918; Canby 1951). The remaining contemporaneous authors were selected by first assembling a list of male and female American authors of the 19th–20th century using Wikipedia⁷ and then selecting a subset of these authors, all of who had a few long works publicly available and spread out over at least twenty years. Also, for the purpose of estimating stable word distributions, shorter works of less than 150 kilobytes in length were excluded. 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).⁸ Elizabeth Stuart Phelps Ward and Amanda Minnie Douglas both began writing about 15 years later in 1866.

The literary prose texts were mainly collected from Project Gutenberg (PG):⁹ this part of the corpus consists of 397 hand-transcribed works; it was supplemented with 158 scanned works from the Internet Archive (IA).¹⁰ In general, we might prefer to choose a hand-transcribed version of a text from Project Gutenberg rather than the possibly more noisy OCR version from the Internet Archive. 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 PG version did not have a time stamp, the IA version was chosen instead if available. On occasion, the OCR versions were manually corrected, but this was determined on an individual basis and through human inspection only.

All data was prepared by manually removing parts that were written at a different time from the main work, along with introductions or

⁷ https://en.wikipedia.org/wiki/Category:19th-century_American_writers – last verified October 2021.

⁸ 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.

⁹ <https://www.gutenberg.org/> – last verified October 2021.

¹⁰ <https://archive.org/> – last verified October 2021.

comments not by the author, such as copyright headers/footers, notes or introductions by editors. Additionally, tables of contents were also removed, as these do not usually follow a normal sentence structure. Klaussner and Vogel (2018a) provides more specific descriptions of the data and its basic pre-processing. The publication date of a text was set by taking the first documented date, e.g. first copyright or publication date, unless a preface clearly stated that the work had been subject to explicit revisions. The issue with dating in this case is that either dating a work too early or too late would distort the results.

The reference language corpus for the current work was assembled by taking an extract from The Corpus of Historical American English (COHA: Davies 2012).¹¹ COHA is a 475-million word corpus that contains samples of American English from 1810–2009, balanced in size, genre and sub-genre in each decade (1000–2500 files each). Depending on the particular type of analysis, different excerpts from the entire data set were used. The corpus contains balanced language samples from fiction, popular magazines, newspapers and non-fiction books, which are again balanced across sub-genre, such as drama and poetry.¹² While the corpus is balanced overall, some years contain proportionally more data from certain genres than others, where we observed strange frequency effects. However, to the best of our knowledge, for our current requirements of providing an approximation to general language usage at the time, this corpus still provided the best option.

METHODS

4

Section 4.1 describes how features were extracted, and is followed by Section 4.2: the statistical models used for the analysis.

¹¹ A free web-based version is accessible on: <https://www.english-corpora.org/coha/> – last verified October 2021.

¹² There is an Excel file with a detailed list of sources available on: <https://www.english-corpora.org/coha/> – last verified October 2021.

We begin by describing the feature extraction adopted by Pennebaker and Stone (2003), interlaced with our own design, where modifications were deemed necessary. As previously mentioned, Pennebaker and Stone (2003) based their analysis on the LIWC system, whose categorization scheme is generally not openly accessible. This renders replication of less objective linguistic variables, such as negative or positive emotion words difficult.¹³

Table 1 only lists examples of non-reflexive uses of pronouns and main tenses, so it is unclear whether reflexive pronouns were included and how complex verb forms also indicating aspect, such as present perfect or future perfect, were treated in their analysis. For extracting 1SG/1PL pronouns in the current work the word was used in conjunction with the part-of-speech tag to identify the correct items, e.g. to avoid uses of *I* that refer to numbering.¹⁴ As our experiments did not show differences between including or excluding reflexive pronouns, this analysis only reports on non-reflexive pronoun types.

Originally, P&S also included what they refer to as “time-related” words, such as *clock*, *hour* and *soon*. One can assume that they would also include temporal adverbs in general like *yesterday* or *today*. These temporal expressions may change the interpretation of regular tenses and could result in shifts between them. However, this may not be a trivial problem, as sometimes the overall tense would be more strongly signaled by the temporal adverb, e.g. examples (1) and (2), whereas in other cases the verb would be the determining factor, as in example (3).

- (1) *She’s there tomorrow.*
- (2) *She’s there today.*
- (3) *She was there today.*

¹³To the best of our knowledge, these words were classified by several different students and can be (indirectly) accessed through the LIWC program. Research papers usually only provide examples rather than exhaustive lists.

¹⁴For all the computations in this work, the statistical programming language R (R. Core Team 2014) and associated packages were used. For POS-tagging the *NLP* (Hornik 2016) and *openNLP* (Hornik 2015) packages were used.

This suggests the need for a more intricate classification system than could be done justice as part of the present work. Here we resort to only using verb tenses to approximate the overall tenses. The main effect of not including temporal adverbs may be a shift from future to present tense counts. In order to approximate tense representation, we adopted the following classification: while POS tags could be used to directly identify some of the simpler tenses, this would not suffice to always correctly determine the difference between the present or present perfect tense usage of *have* and neither could it identify occurrences of the *going-to* future tense, as this is not marked explicitly on *going-to*.¹⁵ To be able to make these distinctions, we used chunk tags to extract verb phrases and then analyzed the combination of tags within to determine the type of tense. In this, several sub-types corresponding to finer shades of difference in meaning are classified into the three main categories (*past/present/future*), as follows. The *present* type includes: simple present, present progressive, and conditional and modal variants, such as *can/could/may go*. The *past* type captures simple past, present perfect, past perfect, past progressive and, as with the *present* type, conditional and modal variants, such as *could have gone*. Finally, the *future* type covers simple future construction, such as *will/shall go* and *going to go*, but also *will have gone*. Finally, we define long-letter sequences as previously, as words whose length is greater than or equal to six letters.

After extracting the relevant features, texts in each corpus were combined by considering the year of publication, thereby reducing each set to one file per year per (author) corpus. Relative frequencies for each feature type were calculated by considering the ratio of the occurrence of the feature and all tokens for the same year. In addition, ordinal variables were created corresponding to year of publication (*year*), age of author at publication of text (*age*) and a categorical variable indicating the author (*A*) of a text.

¹⁵ Still, somehow items such as *I'm going to school* had to be distinguished from *I'm going to go to school*.

This section describes aspects connected to the statistical analysis, i.e. regression models and standardization techniques, before moving on to model assessment.

Temporally-ordered data can be analyzed in different ways, for instance relating a variable to itself at different points in time as part of a “time-series” model or, as in the present case, by considering other variables at the same point in time thereby using an “explanatory model”. Consequently, the prediction of a variable y 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 $t : \{t \in 1, \dots, n\}$, and some error term: $y_t = f(x_{1t}, x_{2t}, \dots, x_{p-1t}, x_{pt}, error)$.

The regression models computed in the following experiments vary with respect to the data set used and whether individual author variation had to be accounted for. The reference corpus (RC) does not contain an *age* variable and is only evaluated with respect to *year* of publication, which serves to check whether a particular variable of interest is likely to have changed in relative frequency over time.

However, when analyzing the literary authors corpus, both *age* and *year* have to be considered as predictors, since the authors will align differently depending on the variable, i.e. James and Twain were not the same age in the same year. Thus, in order to argue for an ageing effect to be present for an individual, it has to (also) be found in a combined model of the authors, clearly outperforming the equivalent year-based model that does not depend on age, but may capture stylistic changes over time instead.

When analyzing different authors at the same time, one may have to resort to random effects models to account for individual variation between authors as shown by Equation (1), where y_{tj} is the response variable for author j at time t , x_{tj} is the individual-specific random effect and A_j is the author-specific random effect; ε_{tj} represents the error term. Similarly, Equation (2) shows the same for the quadratic model, adding predictor $\beta_2 x_{tj}^2$.

$$(1) \quad y_{tj} = \beta_0 + \beta_1 x_{tj} + \beta_2 A_j + \varepsilon_{tj}$$

$$(2) \quad y_{tj} = \beta_0 + \beta_1 x_{tj} + \beta_2 x_{tj}^2 + \beta_3 A_j + \varepsilon_{tj}$$

For fitting linear and normally distributed models, the *nmls* R package was used (Pinheiro *et al.* 2013). Data that was only log-normal was fitted through the *glmmPQL* function in the *MASS* package (Venables and Ripley 2002). In order to preserve similarity with P&S's study, the predictors *age* and *year* were standardized two-ways, one by computing z-scores, i.e. subtracting the mean and dividing by one standard deviation for the simple linear regression models, and also by taking the absolute value of the difference from the mean over the sample for the quadratic models. For correlation analysis, either Pearson correlation coefficient r or Spearman's ρ were used, for normally and non-normally distributed data, respectively.

The decision as to what type of model and correlation measure to use, i.e. parametric or non-parametric, was based on whether the linear model fulfilled all model assumptions: all models were tested for normality, kurtosis, skewness, nonlinear link function (for testing linearity) and heteroscedasticity.¹⁶

EXPERIMENTS

5

This section begins by examining background language change with respect to the six linguistic variables outlined in Table 1. Having considered background language change, Section 5.2 then investigates how these effects can be explicitly modelled in the case of the literary authors. This also allows us to determine to what extent background language may be responsible for effects observed in the individuals.

Background language change

5.1

Examining the change in linguistic variables over time raises the question to what extent these variables were subject to other outside in-

¹⁶ Computed through the *gvlma* package in R (Pena and Slate 2014).

fluences, especially when considering a time span of ~ 40 years or more. To be able to assign meaning to measures of linguistic ageing, a separate analysis of the change in the background language is conducted as part of this section. In general, observed individual effects could be either subsumed by language change or rendered more significant if they happen to be in the opposite direction. Thus, taking background language into account can both lessen and strengthen individual effects.

Table 4 shows correlation results for the reference corpus and both P&S's Disclosure project and Author project. The results for computing simple linear (β) and quadratic (β^2) models are displayed only for the reference corpus alongside the DP as the same model computations were not available for the AP. Our reference corpus shares characteristics with both of P&S's studies in that it covers a similar length of time as the DP (~ 70 years) and years contain multiple individual samples rather than a strict within-subject design. However, it is more comparable to the AP design in that it is genuinely sampled from different time periods, whereas some of the DP's data representing different age groups could have originated from the same time period. For this reason, we aim for a general comparison or replication rather than remaining very close to the original study.

Language change effects can be observed with respect to at least three of the six variables, and this is specifically notable in the case of 1PL pronouns and past tense, where the effect is in the same direction as for the DP, and the case of long-letter sequences, where effects are in the opposite direction for both of P&S's studies.

5.1.1

Change in pronouns

Figure 1 depicts 1SG and 1PL pronouns in the RC over the time span from 1830–1919.¹⁷ As can be observed, 1SG pronouns slightly increase in relative frequency over time. All model parameters in Table 4 show a positive but non-significant trend over time.

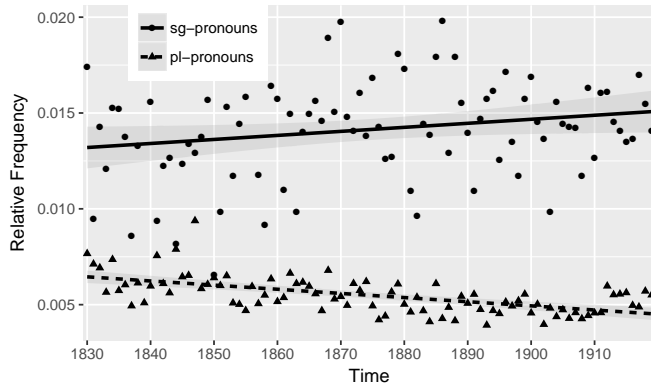
Both P&S's studies have significant, but negative associations for 1SG pronouns over time. 1PL pronouns experience a highly significant decrease in relative frequency over the reference corpus, and

¹⁷ As there were some sampling irregularities in the reference corpus around 1923, the years after 1919 were excluded, resulting in 90 years of data.

Table 4: This table shows correlation analysis (r) and main model coefficients for simple linear (β) and quadratic models (β^2) for both P&S's Disclosure and Author project, and the current 18th-19th century reference corpus. Items marked with '!' signal that linearity assumptions were violated. By default Pearson's r is used, but is replaced by Spearman's ρ for departures from linearity; this is indicated by a superscript ρ . Significance levels are indicated as follows: *: $p \leq 0.05$ /**: $p \leq 0.01$ /***: $p \leq 0.001$

LWC variable	P&S: Disclosure project			P&S: Author project			Reference corpus		
	r	β	β^2	r	r	β	r	β	β^2
Social and identity									
First-person singular	-0.13**	-0.14**	-0.2	-0.26*	0.18 ρ	0.0005!	0.000003!		
First-person plural	-0.12**	-0.13**	0.19**	0.03	-0.60 ρ ***	-0.0006!***	0.0000008!		
Time orientation									
Past-tense verbs	0.04**	-0.16**	0.01	0.08	0.7 ρ ***	0.004!***	-0.000002		
Present-tense verbs	-0.02	0.04*	0.06**	0.09	0.16	0.0003!	0.000005**		
Future-tense verbs	0.00	0.14**	-0.02	0.22*	0.04	0.00001!	0.000002		
Cognitive complexity									
Long-letter seq. (> 6 letters)	0.13**	0.26**	-0.03	0.10	-0.51 ρ ***	-0.02!***	-0.000004		

Figure 1:
Reference
corpus:
1SG and 1PL
pronouns



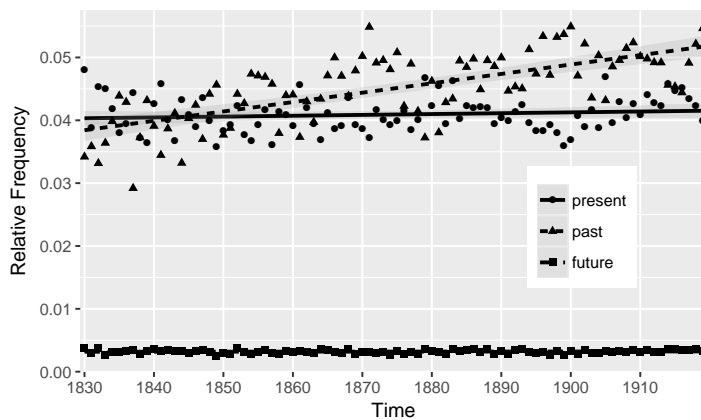
while P&S’s Author project reveals low non-significant correlations, their Disclosure project shares this highly significant downward trend. The linear model results mirror these correlations for both variables. There is less evidence of background language interference in the case of 1SG pronouns, but stronger indications in the case of 1PL pronouns.

5.1.2

Change in tenses

Figure 2 shows relative frequencies for past, present and future tense. Future tense shows little variation over time or at least not at a significant level, while examining Table 4 shows that both P&S’s data sets have a positive association for future tense over time. Present tense appears stable in relative frequency and has a significant positive quadratic trend as can also be observed in P&S’s DP. Past tense in

Figure 2:
Relative
frequencies
of past, present
and future tense



the RC has a highly significant positive correlation (0.7^{***}) and highly significant regression coefficient β , and while r is also positive and significant in P&S's DP, it is reported to have a significant negative linear regression coefficient (−0.16^{**}). Their AP has a non-significant positive correlation for both present and past tense. Both visual and statistical analysis indicate that the tenses, but especially the past tense, underwent change in frequency in background language use for the time period examined, and as with 1PL pronouns could therefore introduce noise into stylistic or ageing analyses.

Change in long-letter sequences

5.1.3

The development of long-letter sequences over the RC is shown in Figure 3. There is a continuous downward trend visible, which is confirmed by both a highly significant correlation coefficient ρ (−0.51^{***}) and a linear regression coefficient β (−0.02^{1***}) in Table 4. Both P&S's DP and AP have positive trends and therefore trends in the opposite direction (r of 0.13^{**} and 0.10 respectively).

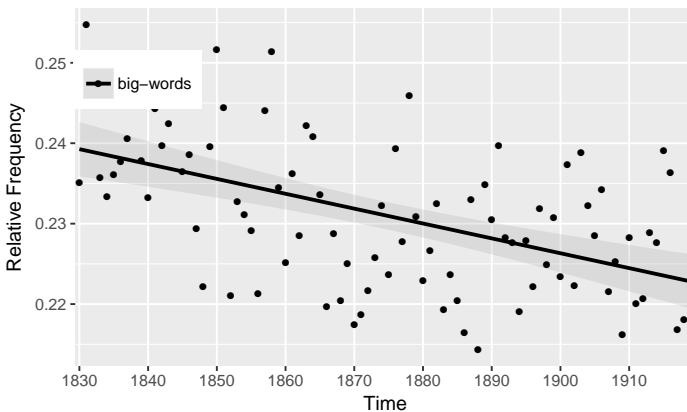


Figure 3:
RC: long-letter
sequences

Discussion

5.1.4

This section has examined six linguistic variables in a continuous section of general language usage that have been hypothesized in the literature to be affected by ageing in individual writers. Pennebaker and Stone (2003) found significant decreases in all their

data sets with respect to 1SG pronouns. For the time frame examined here, no significant trend for 1SG pronouns based on publication year was observed in the reference language. This adds weight to the interpretation of the P&S 1SG effect as being one of ageing. 1PL pronouns were negatively associated with age for the Disclosure study and our reference corpus also showed a highly significant negative trend over time. This suggests that the 1PL effects may not be due to ageing. Pennebaker and Stone's work observed a significant decrease in past tense verbs in the DP, while this variable could be observed to increase in the RC. Present tense was not found to be a likely factor in ageing by P&S, which can be partially confirmed as the relative frequency did not seem to undergo a very pronounced shift. Similarly, there did not appear to be a very strong effect for future tense in our reference corpus, whereas it was found to increase over all of P&S's data sets, possibly implicating this as a real ageing effect. Long-letter sequences are comparable to the past tense situation: Pennebaker and Stone (2003) report a significant increase over their Disclosure project, whereas there is a significant decrease over the background language sample examined here. If their data were subject to similar effects, then this could render the linguistic ageing results more pronounced. This analysis has shown there to exist significant language change in most of the ageing variables examined. To what extent this challenges or amplifies results in the original study is not further examined here. Rather, the next section addresses how these underlying influences can be taken into account when examining linguistic ageing variables in the literary authors corpus, by attempting to estimate the impact of background language change more systematically for the literary authors. We then consider to what extent this underlying change influences interpretation of effects previously only attributed to ageing.

5.2

Estimating impact of language change

In this section, we aim to investigate the ageing hypotheses with respect to the literary authors corpus while controlling for background language influence. For instance, a random effects model as shown

in Equation (3) can be used, taking into account reference language, where ref_{ij} is the relative frequency of the reference language for author j (A_j) at age i and random error ϵ_{ij} . Equation (4) shows the equivalent quadratic model.

$$(3) \quad y_{ij} = \beta_0 + \beta_1 ref_{ij} + \beta_2 Age_{ij} + \beta_3 A_j + \epsilon_{ij}$$

$$(4) \quad y_{ij} = \beta_0 + \beta_1 ref_{ij} + \beta_2 Age_{ij} + \beta_3 Age_{ij}^2 + \beta_4 A_j + \epsilon_{ij}$$

The set of literary authors varied somewhat and for most variables only a subset of authors produced a normal or log-normal fit. For this reason different subsets of the entire data were used to test individual variables' hypotheses.

Table 5 shows the results of computing simple linear random effects models for the six linguistic variables. The first two columns show model coefficients for the age and background language predictors. The third column specifies what model type was used, i.e. normal (N) or log-normal (LN) and the final column lists the respective size of author set. Overall, there is little evidence for either a very strong

Table 5: This table shows the main model coefficients for simple linear regression using random effects models. 'Age.std' and 'Ref.std' refer to standardized age predictor and background change factor respectively. 'Model type' specifies normal (N) or log-normal (LN) setting and '|Authors|' refers to the size of the supporting set. Significance is indicated by: *: $p \leq 0.05$ /**: $p \leq 0.01$ /***: $p \leq 0.001$ /***: $p \leq 0.1$. A '†' on the ageing coefficient indicates that the equivalent model using *year* (of publication) was more significant

LIWC variable	Model coeff.		Model type	Authors
	Age.std	Ref.std		
Social and identity				
First-person singular	0.0008	-0.0002	N	12
First-person plural	0.02	0.07***	LN	15
Time orientation				
Past-tense verbs	0.0008	-0.0002	N	10
Present-tense verbs	0.00009	-0.0004	N	18
Future-tense verbs	-0.0002***†	0.0005	N	20
Cognitive complexity				
Long-letter seq. (> 6 letters)	0.0007	-0.002	LN	21

Figure 4: R output for a *glmmPQL*-based model predicting future tense from reference language and age or year

Fixed effects: response ~ Ref.std + Age.std						
	Value	Std.Error	DF	t-value	p-value	
(Intercept)	0.0029418574	2.019705e-04	335	14.565781	0.0000	
Ref.std	0.0000592255	5.052461e-05	335	1.172211	0.2419	
Age.std	-0.0002297820	5.675550e-05	335	-4.048629	0.0001	
Fixed effects: response ~ Ref.std + Year.std						
	Value	Std.Error	DF	t-value	p-value	
(Intercept)	0.0029857001	0.0001912692	335	15.609939	0.0000	
Ref.std	0.0000626322	0.0000505549	335	1.238895	0.2163	
Year.std	-0.0003035166	0.0000706138	335	-4.298262	0.0000	

influence of background language change or linguistic ageing. The only nearly significant reference language coefficient is 1PL pronouns. Figure 4 presents evidence for some language change influence, i.e. removing the reference language predictor causes the *Year.std* predictor to become significant, while the ageing predictor *Age.std* in the equivalent model does not become more important, indicating that time of publication remains more salient than age of author. The only significant ageing predictor is for future tense; however, considering the equivalent model using *year* (of publication) instead of *age* (at time of publication) renders an even more significant model, calling into question the validity of *age* as a main cause of the observed effect.

Table 6 shows the results for computing quadratic random effect models for the six variables based on Equation (4). Similarly to the simple linear model results, quadratic models also do not yield well fitting models (in terms of significant predictors) for either age or background language predictors. For 1PL pronouns, the reference language predictor is almost significant in the sense of very nearly crossing the threshold for statistical significance at the 95% confidence level, as in the case of the simple linear model in Table 5. Although the ageing predictor for future tense in Table 6 is not significant, the equivalent quadratic year predictor is.

Finally, we turn to the last part of this analysis, namely the question of stylistic differences between authors. For instance, one could consider the question of whether there is likely to be anything particular about Mark Twain's and Henry James' style development compared to the other authors given that these two have received considerable attention from literary scholars. Further, we consider the

Table 6: This table shows the main model coefficients for quadratic regression using random effects models. ‘Age.std²’ and ‘Ref.std’ refer to standardized age predictor and background change factor respectively. ‘Model type’ specifies normal (N) or log-normal (LN) setting and ‘|Authors|’ refers to the size of the supporting set. Significance is indicated by: *: $p \leq 0.05$ / **: $p \leq 0.01$ / ***: $p \leq 0.001$ / †: $p \leq 0.1$. A ‘†’ on the ageing coefficient indicates that the equivalent model using *year* (of publication) was more significant

LIWC variable	Model coeff.		Model type	Authors
	Age.std ²	Ref.std		
Social and identity				
First-person singular	-0.00001	-0.5	N	13
First-person plural	0.03	0.07***	LN	15
Time orientation				
Past-tense verbs	-0.0002	0.1	N	10
Present-tense verbs	-0.00001	-0.02	N	18
Future-tense verbs	-0.000001†	0.2	LN	18
Cognitive complexity				
Long-letter seq. (> 6 letters)	0.005	-0.002	LN	21

specific case of first-person pronouns. Figure 5, Figure 6, Figure 7 and Figure 8 show 1SG and 1PL pronouns for Twain and James alongside some of the other authors in the set, as well as a line representing the average over all authors in the set.¹⁸ Figure 5 shows James and William Dean Howells and Figure 6 shows Twain and Elizabeth Stuart Phelps Ward. For neither Twain nor James does there appear to be a particular development in the form of a trend for 1SG pronouns. Nor is their level of variation around the authors’ average among the highest. As the plots indicate, Howells and Ward show more variation for 1SG pronouns than either Twain or James. Figure 7 and Figure 8 confirm this general impression. For 1PL pronouns, James shows comparatively little variation over time, while Twain’s style displays

¹⁸The ‘aut-ref’ line represents an average over all authors in the set, computed by, for each year, taking the raw frequencies for that year and two years before and after for each author separately, then averaging over all tokens in those years. Given this set of relative frequencies for a feature, the final frequency is given by averaging over all authors for a given year. Hereafter, this is also referred to as ‘author reference corpus’ or ‘ARC’.

Figure 5:
1SG pronouns
for James,
Howells and the
ARC

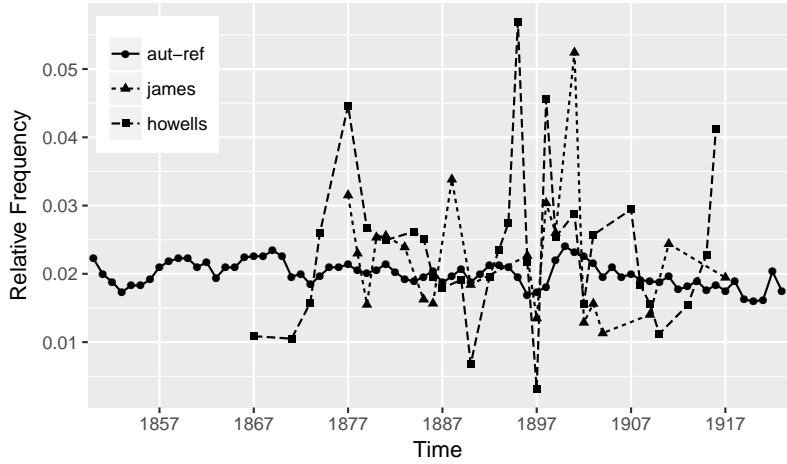
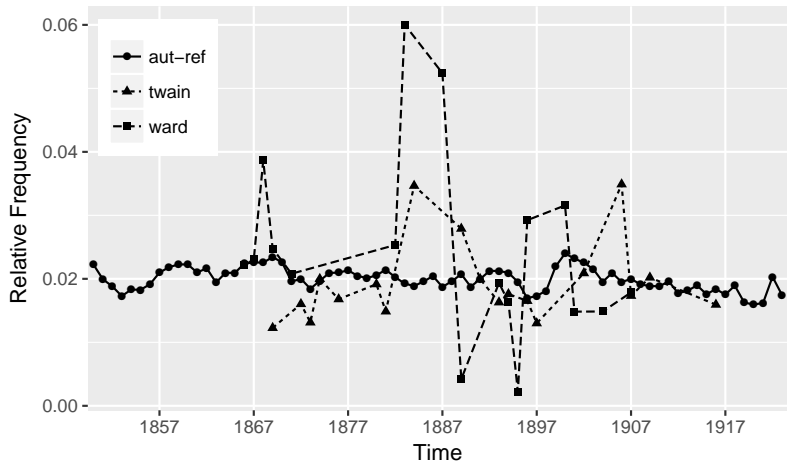


Figure 6:
1SG pronouns
for Twain, Ward
and the ARC



somewhat more variation around the authors' average. However, both authors are not unique in their tendencies. Like James, Alice Brown deviates comparatively little from the average, while Timothy Shay Arthur's relative frequency also increases in his last works similarly to Twain. Thus, there appears to be little evidence that Twain and James are decidedly different from their contemporaries in terms of style change. In the previous section, we identified an effect for 1PL with respect to background language effect, yet overall there is little evidence that there is a systematic influence of age or background language for these literary authors, at least for the variables examined.

Linguistic ageing in literary authors

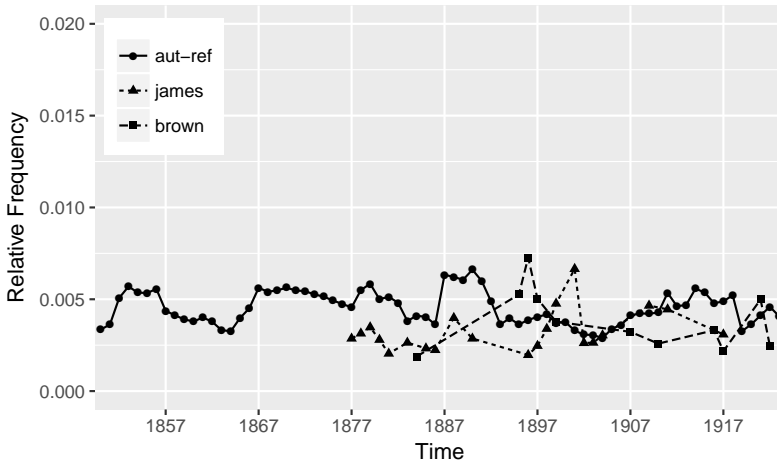


Figure 7:
1PL pronouns for
James, Brown
and the ARC

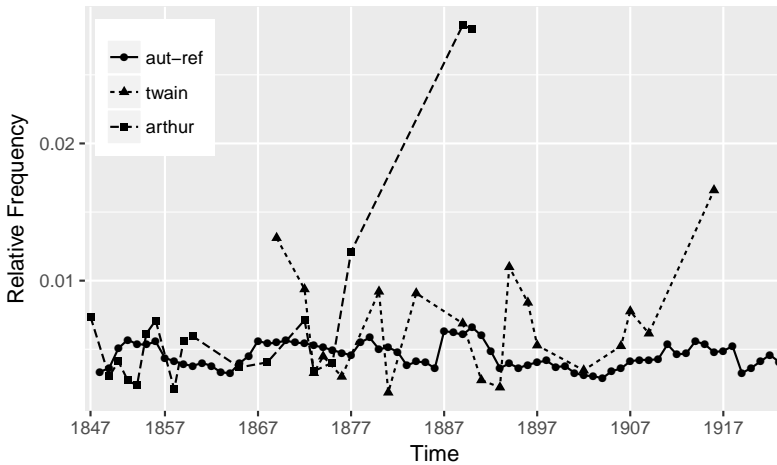


Figure 8:
1PL pronouns for
Twain, Arthur
and the ARC

This could indicate that literary authors have a higher command over their language usage and may be more impervious to outside influences.

DISCUSSION

6

This work has considered aspects of linguistic ageing and how this influences literary authors. In part, the study presented here was a replication of an earlier study by Pennebaker and Stone investigating the

ageing effects in emotional disclosure studies and a corpus of literary authors. Although significant effects were found with respect to pronouns, future and past tense, and long-letter sequences in their study, these results did not replicate with respect to the authors examined here in a unified fashion that would suggest a rise or fall in frequency is actually due to age rather than only stylistic variation of individual authors. The fact that the results of the earlier study could not be replicated may be due to properties of this particular data set, but it could also hint at the possibility of this linguistic ageing effect not existing for professional writers, who could conceivably possess a higher command over their language style than non-professional writers. This would be consistent with P&S's findings insofar as their results for literary authors were also less significant than those for non-professional writers. This does not necessarily challenge the existence of linguistic ageing as a phenomenon, but rather suggests that the variables analyzed here do not provide good proxy measures for it, at least not with respect to literary writers. However, for this analysis no other non-linear models have been examined, something that would have to be done to completely refute the proposed hypotheses with respect to ageing.

The other purpose of this study was to examine these six variables for evidence of language change, and the results indicate significant change in the usage of at least 1PL pronouns, past and present tense verbs, and long-letter sequences. Overall, the models computed above for the literary authors present little evidence that background language (change) had a strong influence on them. However, the models built for 1PL pronouns present some evidence of background language influence, which indicates the necessity to control for it in general. A final result of this analysis was the diversity in the literary authors, which interestingly was not (only) caused by the prominent writers Mark Twain and Henry James. Instead, our analysis suggests that overall they seemed to align well with their contemporaries.

Based on this analysis, it appears that there could be some variation between authors for the six variables examined, possibly indicating stylistic differences with respect to other variables. These differences could be explored in more depth by looking more generally at stylistic change in the literary authors against the backdrop of general language shifts.

CONCLUSION

7

This work has considered to what extent ageing affects language development, examining six linguistic variables that had been reported as significant in the literature. While effects in previous studies were mainly found for non-professional writers, even significant effects confirmed by P&S for literary authors could not be replicated here. This does not necessarily prove an absence of previously identified effects, but calls for additional research to investigate this further. There is strong evidence of background language change for these variables, calling for explicit modelling of this influence, as has been exemplified as part of this work.

ACKNOWLEDGEMENTS

This research is supported by Science Foundation Ireland (SFI) through the CNGL Programme (Grants 12/CE/I2267 and 13/RC/2106) in the ADAPT Centre (<https://www.adaptcentre.ie>).

REFERENCES

- Joseph Warren BEACH (1918), *The method of Henry James*, Yale University Press.
- Helen BIRD, Matthew A. LAMBON RALPH, Karalyn PATTERSON, and John R. HODGES (2000), The rise and fall of frequency and imageability: noun and verb production in semantic dementia, *Brain and Language*, 73(1):17–49, <https://www.sciencedirect.com/science/article/pii/S00939334X00922934>.
- Deborah M. BURKE and Meredith A. SHAFTO (2008), Language and ageing, *Science*, 346:373–443.
- Henry Seidel CANBY (1951), *Turn west, turn east: Mark Twain and Henry James*, Biblio & Tannen Publishers.
- Mark DAVIES (2012), The 400 million word Corpus of Historical American English (1810–2009), in *English Historical Linguistics 2010: Selected Papers from the Sixteenth International Conference on English Historical linguistics (ICeHL 16), Pécs, 23–27 August 2010*, pp. 231–61.

- Richard S. FORSYTH (1999), Stylochronometry with substrings, or: a poet young and old, *Literary and Linguistic Computing*, 14(4):467–478.
- David L. HOOVER (2007), Corpus stylistics, stylometry, and the styles of Henry James, *Style*, 41(2):174–203.
- Kurt HORNIK (2015), *openNLP: Apache OpenNLP tools interface*, <https://CRAN.R-project.org/package=openNLP>, R package version 0.2-5.
- Kurt HORNIK (2016), *NLP: natural language processing infrastructure*, <https://CRAN.R-project.org/package=NLP>, R package version 0.1-9.
- Susan KEMPER, Lydia H. GREINER, Janet G. MARQUIS, Katherine PRENOVOST, and Tracy L. MITZNER (2001), Language decline across the life span: findings from the nun study, *Psychology and Aging*, 16(2):227–239.
- Carmen KLAUSSNER and Carl VOGEL (2018a), A diachronic corpus for literary style analysis, in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA).
- Carmen KLAUSSNER and Carl VOGEL (2018b), Temporal predictive regression models for linguistic style analysis, *Journal of Language Modelling*, 6(1):175–222.
- Xuan LE, Ian LANCASHIRE, Graeme HIRST, and Regina JOKEL (2011), Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists, *Literary and Linguistic Computing*, 26(4):435–461.
- Erez LIEBERMAN, Jean-Baptiste MICHEL, Joe JACKSON, Tina TANG, and Martin A. NOWAK (2007), Quantifying the evolutionary dynamics of language, *Nature*, 449(7163):713.
- Edsel A. PENA and Elizabeth H. SLATE (2014), *gvlma: global validation of linear models assumptions*, <http://CRAN.R-project.org/package=gvlma>, R package version 1.0.0.2 (last verified: 24.08.2015).
- James W. PENNEBAKER and Lori D. STONE (2003), Words of wisdom: language use over the life span, *Journal of Personality and Social Psychology*, 85(2):291–231.
- Jose PINHEIRO, Douglas BATES, Saikat DEBROY, Deepayan SARKAR, and R. CORE TEAM (2013), *nlme: linear and nonlinear mixed effects models*, R package version 3.1-113.
- R. CORE TEAM (2014), *R: a language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, <http://www.r-project.org>.
- Sanja ŠTAJNER and Ruslan MITKOV (2011), Diachronic stylistic changes in British and American varieties of 20th-century written English language, in *Proceedings of the Workshop on Language Technologies for Digital Humanities and*

Linguistic ageing in literary authors

Cultural Heritage at RANLP, pp. 78–85,
<https://www.aclweb.org/anthology/W11-4112/>.

Constantina STAMOU (2007), Stylochronometry: stylistic development, sequence of composition, and relative dating, *Literary and Linguistic Computing*, 23(2):181–199.

W. N. VENABLES and B. D. RIPLEY (2002), *Modern applied statistics with S*, Springer, fourth edition, <http://www.stats.ox.ac.uk/pub/MASS4>.

Carmen Klaussner

© 0000-0002-5469-2167
klaussnc@tcd.ie

Carl Vogel

vogel@tcd.ie


Arnab Bhattacharya

© 0000-0003-2828-7691
bhattaca@tcd.ie

School of Computer Science
and Statistics, O'Reilly Institute,
Trinity College Dublin, Ireland

Carmen Klaussner, Carl Vogel, and Arnab Bhattacharya (2021), *Aspects of linguistic ageing in literary authors across time*, *Journal of Language Modelling*, 9(2):195–223

doi <https://dx.doi.org/10.15398/jlm.v9i2.270>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.
©  <http://creativecommons.org/licenses/by/4.0/>