Elements of Style Change

Carmen Klaussner

Thesis submitted for the Degree of Doctor of Philosophy

School of Computer Science & Statistics

Trinity College

University of Dublin

September 2017

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. Wherever there is published or unpublished work included, it is duly acknowledged in the text.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Summary

This thesis considers aspects of stylistic change over time with respect to a corpus of literary authors and corresponding background change for the same period of time focusing on features that appear in all time instances examined. Within this, it addresses three different dimensions: possible effects of ageing on specific linguistic variables, methods for detection of stylistic changes and methods for detection and interpretation of more sudden changes in frequency.

More specifically, Chapter 1 introduces the field, motivates this research and states the research questions addressed as part of this thesis.

Chapter 2 outlines related work in the field and identifies key studies that this research builds on.

Chapter 3 describes the two main data sets used in this study: an American literary authors' corpus spanning published works from 1847–1923 and a reference corpus for American English covering the years 1830–1929. The chapter begins by providing some background information about Henry James and Mark Twain. Both were prolific and influential authors, who have frequently been targeted by literary scholars and for whom there exist findings that would suggest them to be likely candidates for stylistic change. Hereafter, the literary authors' corpus is introduced more formally, noting collaborations and relationships between the authors within. Then, the reference corpus is described, followed by data preparation and aspects of part-of-speech tagging.

Chapter 4 investigates effects of ageing onto specific linguistic variables reported as significant in the literature. Their effect is first considered with respect to the background language change at the same time, finding that the majority of these ageing features underwent general change in usage that could have affected the interpretation of effects in the individual writers. The final part addresses the question of how to include background language influence in a linear model for the literary authors. In addition, James and Twain are compared to the remaining authors in the corpus with respect to some of the ageing variables. No specific ageing effects can be found in the literary authors' language with respect to the variables suggested in the literature.

Chapter 5 then continues the analysis of style over time in literary authors with respect to general language change by introducing methods to detect salient features through a prediction task. It discusses some of the discovered features and provides a model for taking reference

language change into account when modelling stylistic changes in individual authors.

Finally, Chapter 6 addresses the question of how the features that appear in all time instances and that have primarily been focused on here, relate and are influenced by other less regularly appearing items. The type of change targeted in this is sudden and sustained change in frequency rendering influence of other types more likely. The initial detection analysis indicated that temporal expressions in news data may be interesting to consider in this context. The results and comparisons to other genre suggest that there may indeed have been clusters of irregular words that affected the frequency of the more regularly appearing expressions.

Chapter 7 summarises the results with respect to the research questions and concludes this thesis.

Related Publications

- Klaussner, Carmen and Vogel, Carl. Temporal Predictive Regression Models for Linguistic Style Analysis. *Journal of Language Modeling*, 6(1):175–222, 2018b
- Klaussner, Carmen and Vogel, Carl. A Diachronic Corpus for Literary Style Analysis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7–12 2018a. European Language Resources Association (ELRA)
- Klaussner, Carmen; Vogel, Carl, and Bhattacharya, Arnab. Detecting Linguistic Change Based on Word Co-occurrence Patterns. In *HistoInformatics@ CIKM*, pages 14–21, 2017
- Klaussner, Carmen and Vogel, Carl. Revisiting Hypotheses on Linguistic Ageing in Literary Careers. Paper presented at the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing) Budapest, Hungary, 2017
- Klaussner, Carmen and Vogel, Carl. Stylochronometry: Timeline Prediction in Stylometric Analysis. In *Research and Development in Intelligent Systems XXXII*, pages 91–106. Springer, 2015

Acknowledgement

Foremost, I would like to thank my supervisor Carl Vogel for his excellent supervision over the past four years and especially for indulging and sharing my passion for temporal style analysis. To borrow from a dedication of PG Wodehouse's: *"without [his] never-failing sympathy and encouragement this [thesis] would [probably] have been finished in half the time."*[1]

My sincere thanks also go to Arnab Bhattacharya for lending statistical support and for great collaboration. Further, I would like to thank Martin Emms and Myra O'Regan, my confirmation examiners, who gave very constructive criticism that greatly helped to improve the earlier outline of this thesis. Also, I would like to thank my previous supervisors and/or co-authors Desislava Zhekova, John Bateman (during my BA) and John Nerbonne and Çağrı Çöltekin (during my MA) for not only being very inspired but also very inspiring researchers.

I would like to thank my esteemed colleagues that gave academic and social support during the past four years and in particular: Liliana Mamani Sanchez, Maria Koutsombogera, Justine Reverdy, Akira Hayakawa, Erwan Moreau, Gerard Lynch, Arun Jayapal, Kevin Doherty, Shane Sheehan, Alfredo Maldonado Guerra and Annalina Caputo.

My sincere thanks also go to Scott Ahern and Karita Cullen for their help and support during some difficult months.

Further, I would like to thank the SCSS help in Trinity for their extremely fast response in fixing computer stuff, and also the library and administrative staff for their continued support throughout this time.

Finally, I would like to thank my non-academic friends for not only their support and understanding during this time, but also for reminding me that there is a life outside the Ph.D worth living.

And lastly but most importantly I would like to thank my family for all their support throughout my life, but in particular during the last few years providing me with what I needed to make it this far.

[1] Inspired by a dedication of Wodehouse to his daughter Leonora in *The Heart of a Goof (1926)*.

Prologue

One fine winter's day when Piglet was brushing away the snow in front of his house, he happended to look up, and there was Winnie-the-Pooh. Pooh was walking round and round in a circle, thinking of something else, and when Piglet called to him, he just went on walking.
'Hallo!' said Piglet, 'what are *you* doing?' [...]
'Tracking something,' said Winnie-the-Pooh very mysteriously. [...]
'Tracks,' said Piglet. 'Paw-marks,' [...] Do you think it's a–a – a Woozle?' [...]
'You can never tell with paw-marks.'
With these few words he went on tracking, and Piglet [...] ran after him. Winnie-the-Pooh had come to a sudden stop, and was bending over the tracks in a puzzled sort of way. [...]
'It's a very funny thing,' said Bear, 'but there seem to be *two* animals now [...]'
'Would you mind coming with me, Piglet, in case they turn out to be Hostile Animals?' [...]
So off they went together. [...]
'The tracks!' said Pooh.'*A third animal has joined the other two!*' [...]
And so it seemed to be. There were tracks; crossing over each other here, getting muddled every now and then [...].
'*What's that?*
Pooh looked up at the sky, and then, as he heard the wistle again, he looked up into the branches of a big oak-tree, and then he saw a friend of his. [...]
Christopher Robin came slowly down his tree. 'Silly old Bear,' he said,'what *were* you doing? First you went round the spinney twice by yourself, and then Piglet ran after you and you went round again together [...]'
'Wait a moment,' said Winnie-the-Pooh, holding up his paw [...] [t]hen he fitted his paw into one of the tracks ... and then he scratched his nose twice, and stood up.
'Yes,' said Winnie-the-Pooh.
'I see now,' said Winnie-the-Pooh.
'I have been Foolish and Deluded,' said he, 'and I am a Bear of No Brain at All.' [...]
'Anyhow,' he said, 'it is nearly Luncheon Time.'

From *Winnie-the-Pooh* by Milne and Shepard [2013]

Contents

List of Tables

List of Figures

List of Abbreviations

Table 1 – Literary Authors' Abbreviations

ID	Abbreviation	Author
1	ab	Alice Brown
2	amd	Amanda Minnie Douglas
3	cdw	Charles Dudley Warner
4	cfw	Constance Fenimore Woolson
5	es	Edgar Saltus
6	espw	Elizabeth Stuart Phelps Ward
7	ew	Edith Wharton
8	fmc	Francis Marion Crawford
9	ga	Gertrude Atherton
10	haj	Horatio Alger jr
11	hbs	Harriet Beecher Stowe
12	hj	Henry James
13	hm	Harold McGrath
14	lma	Louisa May Alcott
15	mh	Marion Harland
16	mt	Mark Twain
17	rwc	Robert W. Chambers
18	sw	Susan Warner
19	tsa	Timothy Shay Arthur
20	us	Upton Sinclair
21	wdh	William Dean Howells
22	wta	William Taylor Adams

Table 2 – Part-of-speech tags used in the Penn Treebank Project.

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Chapter 1

Introduction

Language is subject to *constant* change. There are at least two dimensions to this change: the change in a specific language variety, such as American or Irish English and the change in individuals whose language use is subject to this variety and its change over time.

1.1 Motivation

This thesis is about change in style and more particularly focused on individuals that have made writing their profession, in this case literary authors. Although the field of 'stylometry' and by extension 'authorship attribution' has been witness to considerable advances in methods to characterise writing style and tell different authors apart, methods designed specifically to analyse style change have been somewhat neglected. This is in spite of the fact that usually works of an author's entire lifespan are considered for synchronic analyses, disregarding possible systematic changes he or she might have undergone. In a sense, synchronic studies are thus reduced to detecting those items for which an author did not show any or little change over time, discarding style features that may reflect the development of the author's style in favour of those that reflect what is static about it. However, simply analysing the style of an individual author or even a group of writers will not be able to identify stylistic changes that are truly individual. This can only be achieved by considering the general background language change at the same time thus providing a clearer picture of what is stable and what is changing in a writer's linguistic style.

Consequently, this thesis analyses literary style change taking into account the underlying general language change aspect and endeavouring to quantify what aspects are subject to individual rather than general language change. For this purpose, a corpus containing twenty-two literary authors[1] and and a temporally matched reference corpus[2] were assembled. Specifically, the corpus features ten female and twelve male American literary authors; the latter group

[1] Hereafter also referred to as: 'literary authors' corpus' or 'LAC'.
[2] Hereafter also simply referred to as: 'reference corpus' or 'RC'.

includes Mark Twain and Henry James. Both authors have received considerable attention through stylistic and literary analyses with some of these analysing temporal changes, suggesting that these two authors in particular might be interesting to examine over time. Therefore, a side-question investigated as part of this research is whether this particular attention proves to be justified in the form of a very pronounced change in style that is markedly different from contemporaneous writers.

The primary focus of this thesis, however, is the development of methods for the analysis of authorial style over time, in particular with respect to how ageing affects language over the lifespan, what type of linguistic features are particularly changeable and how different features interact with each other. Rather than providing an exhaustive treatment of stylistic change with respect to the examined corpora, the purpose is to extend the toolkit for analysis and interpretation of diachronic style. In this, the models introduced are seen as "simplification[s] or approximation[s] of reality and hence will not reflect all of reality" [Burnham and Anderson, 2003, p.20]. The objective is not to model the truth, but rather to try to find models approximating the data fairly well in order to learn something about the data [Burnham and Anderson, 2003].

This research is focused on a specific subset of features with the defining property that they appear consistently over all time instances examined. Hereafter, these are referred to as 'constant' features. One of the questions that naturally emerges when considering a special subset of a larger group of features is what justifies examining this set as a special case. Constancy in general should be regarded as a continuum rather than a binary classification of a feature into *being always there* vs. *being never there*. At one end of the spectrum, there are *hapax legomena* or the words that only appear once in the entire corpus with features, whereas at the other end there reside the constant features that appear in all texts. Features that are very infrequent are arguably closer in interpretation to *hapax* items and features that are almost constant could be considered more similar to fully constant features. Fully constant features themselves could be divided into different categories: function words and more frequent open-class words. Function words are used to construct correct grammar in the language and with respect to this group, authors' styles are more likely to vary along the frequency spectrum rather than the constancy one. The other main category is open-class words, such as verbs or adjectives that are still frequent enough to be constant over an author's works. However, there might be more pronounced differences between different authors for these features.

What makes constant features interesting is the fact that as a group they carry intrinsic importance for the author who uses them. Other features that are only partially constant might not occur for other, non-stylistic reasons, such as changes in topic. In contrast, relative frequency changes in constant features could be more readily attributed to a change in style. Apart from these more psycho-linguistic motivations, constant features may display more statistical stability through their regular distributions resulting in more reliable model estimations. For these reasons, this particular subset of features is considered more particularly as part of this research.

1.2 Research Question

This thesis seeks to contribute answers to the following main question, that is divided into four components:

Are there reliable methods that lead one to quantify what is stable and what is changing in an author's style over time in relation to background language in the same period?

1. *What aspects of linguistic change are likely due to ageing ?*
2. *What features are most characteristic of literary change ?*
3. *What change is sudden as opposed to gradual and what are possible interpretations for this?*
4. *Given that Henry James and Mark Twain have been in the focus of literary analyses, is there anything different about their style development?*

1.3 Contribution

Concerning Q1, based on the analysis conducted in Chapter 4, there has been no evidence of systematic linguistic ageing in literary authors. Chapter 5 considers Q2 and finds that the features most characteristic of style change in literary authors are character and syntactic features, in particular character trigrams and tetragrams and part-of-speech bigrams. With respect to Q3, in Chapter 6 temporal expressions emerge as interesting candidates as they display very abrupt change in news language and to some extent also in fiction. While there is evidence of clusters of non-constant words affecting temporal expressions' frequency for the news data, the source of change in literary style may be more subtle. Finally, as for Q4, having considered James' and Twain's style change with respect to ageing, general stylistic change and sudden style change, there is no evidence that their style varies more over time or in very different ways from their contemporaries.

The specific contribution of this thesis is the introduction of methods for temporal stylistic analysis, thereby exemplifying their usefulness for a selection of features. This investigation into temporal linguistic change is by no means exhaustive, but meant to *open the gates* to further explorations of diachronic style.

1.4 Structure

This work is structured as follows: Chapter 2 describes previous work in the field. Chapter 3 presents the two data sets used for this research. Chapter 4 considers aspects of linguistic ageing in literary authors in particular looking at the confounding factor of background language

change. It then presents models explicitly accounting for general language change in the individual author's case. Chapter 5 extends this to detection of general aspects of style change and modelling of background language effects. Chapter 6 relates abrupt changes in constant features to occurrence patterns of non-constant features. Chapter 7 summarises the findings of this thesis, i.e. although there is no evidence for ageing in the literary authors, there is evidence for individual literary style change and also for underlying language change influencing individual writers.

Chapter 2

Literature Review

2.1 Introduction

The style of an author is often referred to as his or her stylistic *fingerprint* and while this is instinctively understood this analogy is scarcely fitting as authorial style has proved to be far more intangible and elusive than this analogy would suggest. In spite of over a century's worth of authorship attribution research and the revolution of fast and efficient computing methods, it has not yet been determined how to tell authorship accurately and it may very well never be possible to do so [Burrows, 1992]. Another central difference to the human fingerprint is that an author's style cannot be determined by the sole consideration of his or her writings. An author's style only takes shape through the comparison with other contemporaneous authors, the exact selection of which determining how close one comes to the *actual* fingerprint. However, in absence of methods to determine stylistic uniqueness without considering other writings, one ultimately always considers a task of authorship attribution which begs the question: "If you can tell authors apart, have you learned anything about them?" [Craig, 1999, p.103]. Thus, in order to validate one's findings, one needs to draw on outside sources, unrelated to the techniques that gave rise to the discoveries, to first confirm the results and thereby also the associated technique. In the case of literary style, these outside sources essentially materialise as literary scholars deciding what words (or other elements) a particular author favours or avoids with respect to other contemporary writers. However, it must be noted that such scholars may be more susceptible to bias in this word selection than automated methods. By extension, this also includes research into the lives and psychologies of individual authors that can be related to through research into the psychological meaning of words [Chung and Pennebaker, 2007].

Yet, regardless of the separate issue of establishing accurate detection methods, one may consider all types of stylistic analyses to hold an intrinsic flaw: although most authors compose their canon over the period of 20–40 years, this aspect is presently not reflected in stylistic analyses. This neglect poses an issue for synchronic style analyses in two different ways. The first being that unless style is found to be invariant for an author and does not change with age and

experience, temporality can be a confounding factor in stylometry and authorship attribution [Daelemans, 2013]. However, if style is not invariant, in the best of the cases the (synchronic) stylistic analysis might merely select those elements *stable* over time as this renders them also stable and consistent over the author's corpus, heedlessly discarding those elements that show temporal variation and incidentally style *development*. The second issue for both synchronic and diachronic analyses presents itself through the phenomenon of general language change that pervades all contemporaneous authors' styles. Consequently, even in the cases where an author's truly individual style shows little variability over time, his overall style would still be subject to the shift affecting all language possibly causing fluctuations that might distort the results.

While temporally invariant words (or other elements) are not uninteresting, elements of style that the author found (even subconsciously) important enough to change over his or her creative lifespan, might hold the key to unlock greater mysteries of style as well as possibly novel aspects of the author's personality. Thus, while also being of great importance for synchronic analyses of style, temporal stylistics (or 'stylochronometry') presents a study in its own right, being the investigation into what is stable and what is changing over an author's life time taking into account general linguistic shifts. In the following, both aspects of general and diachronic style analysis are discussed, specifically the remainder of this chapter is structured as follows: the part immediately following sets the scene by discussing how statistical methods first came to be used in the field of stylometry, thereafter aspects of general style analysis are discussed in Section 2.2, followed by feature analysis in section 2.3. Having dealt with general aspects of stylometry, more specific topics in language change are addressed, i.e. general language change in Section 2.4, followed by language decline and linguistic ageing in Section 2.5, Section 2.6 discusses common methods for temporal analysis and finally Section 2.7 deals with diachronic linguistic style in particular. Section 2.8 concludes this chapter.

The Rise of Statistical Methods The analysis of the style of an author, as for instance discerning what words the author favours or avoids, can be applied in different ways: there are the traditional methods that dominated the scene until well into the 20$^{\text{th}}$ century, which essentially involve literary scholars determining which elements were distinctive of an author compared to other writers at the time. Quantitative attempts at the task, such as American physicist, Thomas Mendenhall's studies of word-length distribution existed, but had not yet met with more widespread acceptance [Mendenhall, 1887, 1901]. Mendenhall had considered differences between writers, such as Charles Dickens and William Thackeray by looking at word length histograms, extending English logician Augustus de Morgan's original suggestion that average word length could be an indicator of authorship [Mendenhall, 1887]. Mendenhall found that two histograms or 'characteristic curves of composition' for different works of the same author were somewhat irregular when only based on 1000 words, whereas two 100,000 words based curves were practically identical. This led him to conclude that in order to show

that the method was sound enough to be used in cases of disputed authorship, it would need to be applied repeatedly and to different authors, i.e. for each author, several 100,000 word curves would have to be compared and found to be identical in the same-author case, while being sufficiently dissimilar in the different-author case. In 1901, Mendenhall conducted a second study, where he attempted to settle the question of Shakespeare's authorship, in particular the question of whether Francis Bacon had authored any of Shakespeare's plays, poems or sonnets [Mendenhall, 1901]. Although Bacon's curve proved to be quite dissimilar to the one of Shakespeare, Christopher Marlowe's curve agreed with one of Shakespeare as much as Shakespeare's curves agreed with themselves. Since then word length has been argued to be a better indicator of register or genre rather than authorship attribution in general [Oakes, 2014, p.5]. The merit of different feature types is discussed in more detail in section 2.3.1.

Among related statistical studies following this early attempt was the influential work by George Kingsley Zipf in 1932 establishing 'Zipf's law' on word frequency distributions in natural language corpora, stating that the frequency of any word is inversely proportional to its rank in the frequency table [Zipf, 1932]. This means that for any natural language corpus, given a most frequent word w, the second-most frequent word accounts for half of w's occurrence, the third-most frequent word accounts for a third of w's occurrence and so on. This finding does also have implications for stylometry, as this is to some extent the study of an author's word distributions, especially the most frequent words.

For several decades, there was no considerable advancement in authorship attribution studies until well into the second half of the 20^{th} century, which marked the emergence of what was to become one of the most famous and influential studies into disputed authorship. In 1964, the two American statisticians Frederick Mosteller and David Wallace set out to use word frequencies to investigate the mystery of the authorship of 'The Federalist Papers' [Mosteller and Wallace, 2008].

During the years of 1787–1788, Alexander Hamilton, James Madison and John Jay wrote the *Federalist Papers* in an endeavour to persuade the citizens of New York to ratify the constitution. The question of authorship arose because originally all articles had been published under the pseudonym of 'Publius' and for 12 papers both Hamilton and Madison later claimed authorship. Even considering additional factors and outside accounts could not settle the dispute satisfactorily. Consequently, Mosteller and Wallace conducted an extensive study as to who wrote the 12 disputed papers, which to complicate matters all had to be attributed individually. Analysis using ordinary style characteristics, such as average sentence length did not yield suitable variables for discrimination between the two authors, which led them to word count analysis. The authors preliminary concluded that one single word or a few words would not provide a satisfactory basis for reliable authorship identification, but that many words in unison were needed to create an "overwhelming" evidence, that no clue on its own would be able to provide likewise [Mosteller and Wallace, 2008, p.10]. They embarked on the laborious task of examining word distributions in the search of words with good discriminatory power.

High frequency words (mostly function words) seemed to provide better discriminators, being both frequent and less subject to contextual influence. The authors conducted a variety of studies, at the heart of which lay a Bayesian likelihood estimation intended to provide an approximation of the prior distributions that were needed to determine conditional/posterior probabilities. Given a vector of word frequencies with density $f_1(x)$ for Hamilton, and $f_2(x)$ for Madison, the likelihood ratio is [Watson, 1966]:

$$\frac{f_1(x)}{f_2(x)} \textit{ and prior probabilities}: \pi_1, \pi_2 \Rightarrow \frac{f_1(x)\pi_1}{f_2(x)\pi_2} \ (\textit{final odds}) \tag{2.1}$$

A paper could then clearly be attributed to Hamilton, if $f_1(x)\pi_1 > f_2(x)\pi_2$ and to Madison if $f_1(x)\pi_1 < f_2(x)\pi_2$. After additional analyses, the authors were able to attribute all 12 papers to Madison and for each paper $\frac{f_2(x)}{f_1(x)}$ was so large as to render any conceivable $\frac{\pi_1}{\pi_2}$ insignificant [Mosteller and Wallace, 2008]. Mosteller and Wallace's work marked the departure point for non-traditional authorship attribution studies to be established more firmly alongside the then predominantly human-expert-based techniques [Stamatatos, 2009].

The rise of the internet in the late 1990s, as well as the increasing availability of electronic texts helped authorship attribution to gather some momentum and be considered alongside other text categorisation tasks, targeted not only with computer-assisted but also computer-based methods [Stamatatos, 2009]. Although new technical advances would allow for more intricate machine-learning methods to be applied to the task of authorship studies, the field tends to be dominated by methods in the statistical realm that are easier interpretable than most of the more obscure machine learning techniques, such as 'deep neural networks'. Interpretability of results and hence the methods that produced them is key in studies of literary style analysis, but also, for instance in forensic linguistic studies, where humanly understandable evidence is of the essence [Clark, 2011].

2.2 Style Analysis

As style is a rather intangible entity with no clear gold standard to evaluate findings, there is a strong need for consistency and intuition about the methods used for detection, analysis and evaluation. In absence of reliable methods for this, one might fall prey to the vicissitudes of statistical optimisation, as for instance failure to reject models based only on mining random correlations in the data. Consequently, there should be an emphasis on using methods for both identification and evaluation of stylistic markers with a strong underlying intuition about them. Especially when using statistical methods, one may happen upon significant results by simply "bombarding [one's] texts with so many tests that some would be bound to show some kind of distinction or identity between baseline text *a* and sample set *b*, simply by the law of averages, regardless of whether the distinction is true or false" [Rudman, 2003, p.28]. Therefore, this section focuses predominantly on more interpretable methods in the field, in particular, sec-

tion 2.2.1 begins by discussing authorship attribution studies; section 2.2.2 continues with the task of document dating and section 2.2.3 considers aspects of stylistic homogeneity.

2.2.1 Authorship Attribution

Studies in the field of style analysis or stylometry focus on different sub-tasks, such as authorship attribution; i.e. given an unknown document and several candidate authors, the task is to decide which candidate is most likely to have authored the document. This problem can be studied in a closed-class and an open-class scenario. The former assumes that the true author is among the set of candidates rendering the task of determining who authored the document in question simpler than in the open-class variant, where the candidate set may or may not contain the true author. Some of the more commonly encountered problems in the open-class attribution scenario is that there could be thousands of possible candidate authors, the text of disputed authorship might not be by any of them or that the text might be of limited length [Koppel et al., 2011]. The exact choice of candidate set is going to be highly significant in both the open-class and the closed-class setting.

One of the earlier, most influential methods in the field has been introduced by John Burrows, namely Burrows' 'Delta' and can be applied in both open-class and closed-class scenarios. Delta was specifically designed for authorship attribution, seeking the most likely authorial candidate for a given document from a set of authors based on differences between z-scores of high-frequency items [Burrows, 2002]. In particular, Delta is computed by first constructing a 'main author' set and then calculating mean μ and standard deviation σ for each word's relative frequency to estimate the average use across different authors in that set. After constructing standardised counts for individual authors in the set and test pieces not in that set, z-scores are computed for each feature separately for all texts by using the main author set's μ and σ computed earlier. To then obtain the distance between a particular author and a test piece, first differences between z-scores are computed for each feature, which are then combined into Delta by averaging over the absolute differences. Delta is defined as "the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text" [Burrows, 2002, p.271]. The technique does not rely on the frequencies directly, but transfers them to a *neutral* scale that allows for a balanced view of the features considered. Although Delta is primarily used for the comparison of individual authors in the main set and a specific test piece outside this set, where the lowest distance indicates the closest fit, one can also compute distributions over all Delta scores. Computing the normal distribution over all the Delta scores – all the individual comparisons – allows one to see whether a particular value diverges a lot from the mean of all differences. If this is the case, then the author considered and the test piece are unusually close and there's no other close competitor (in the set) – this can be further quantified through the z-distribution, hence the values giving rise to the distribution are referred to as 'Delta z-scores'.

Delta, while rather dependent on a strong similarity between the sample texts of one author, is simple to compute and has a way of quantifying how unusual a Delta score is in comparison to other authors' scores in the set.

The 'Zeta' and 'Iota' measures, rather than attempting to assign one text to authorship of one out of many authors, could be considered a complementary measure as they focus on determining which of many texts are most likely to belong to a particular author [Burrows, 2007]. For both measures, a word frequency list is created for a piece of text by a primary author *A*. The sample is then divided into several sections. The crucial step is then to record which words occur in what proportion of those sections as well as how many times in the counter-set *B*, which can consist of only one other author's works or in fact a set of different authors. For Zeta, one retains the words occurring in at least three of 'the five' sections of *A* and then removes those words exceeding a particular frequency in the case of two-author comparisons or in the case of many-author tests, words that occur in most of the authors' samples are removed. For Iota, all the words that appear in just one or two of the five sections are retained, then removing all the words also appearing in the counter-set *B* for two-author tests and for many-author tests removing those words occurring in more than half of the others. The Zeta or Iota score is then computed by taking the total frequency (per 1000 word tokens) of all words that remain after each individual procedure and the higher the score of a text, the more likely it is that author *A* wrote it. Thus, Zeta and Iota target the lower to middle range frequency spectrum, disregarding the very frequent words that the Delta analysis requires. There is evidence to suggest that Zeta and Iota appear to capture genuine authorial idiosyncrasies [Hoover, 2007]. In addition, both Zeta and Iota are remarkably effective in attributing poems as short as 1,000 words [Hoover, 2008]. With respect to Delta there is more recent evidence to suggest that it is particularly suited to prose in English and German, performing less well for agglutinative languages, such as Polish or Latin with results sometimes being improved by not considering the very most frequent words, which are usually at the heart of the analysis [Rybicki and Eder, 2011].

Automated techniques in the domain of authorship attribution can be separated according to two main dimensions: 'similarity-based' methods and 'machine-learning' methods. The methods introduced by Burrows discussed above belong to the type of similarity-based methods, where a metric is used to computationally measure how similar two documents are to each other, where one document is of unknown authorship and the other by a candidate author (his known writing is considered as one document). Based on this similarity score, one determines the most likely author for the anonymous document out of the set of candidates. In machine-learning settings, on the other hand, pieces of known writings form distinct items as part of a training set used to build a classifier that can then be used to attribute anonymous documents [Koppel et al., 2012].

Author Verification Koppel et al. [2007] use a machine-learning based approach to under-

take the task of authorship verification, a sub-task of authorship attribution where given an author and some sample writings, one is asked to determine whether the author wrote a second sample of texts or not, and which is thus part of the open-class problem already discussed above. Koppel et al. [2007] show that the task of deciding whether an author has written a particular text can be accurately determined by iteratively removing the set of most discriminative features from the learning process, a procedure they term 'unmasking'. In particular, given two (long) texts A and X, they train a classifier to distinguish these two texts. Using cross-validation results, they iteratively remove the features best for distinguishing between the two texts from the model and examine how fast accuracy in distinguishing the two texts drops. Differences between texts by the same author are likely to be only reflected in a relatively small number or features causing the model accuracy to drop a lot faster and more dramatically than when the texts were not written by the same person. The authors report 95.7% accuracy with all but one of the same-author pairs being accurately classified and 181 out of 189 different-author pairs being correctly classified. In a second experiment, they use negative-examples or 'impostors' to aid classification. This proves not to be as accurate as the unmasking method, as the method often incorrectly concludes that a given author wrote a book, whereas in the opposite scenario of concluding that an author did not write a particular book it is almost always correct. The authors conclude that although the impostor method is not as accurate as the unmasking method, the former could be used to augment the latter as part of a multiple-classifier setting. However, creating a representative set of impostors is introducing another obstacle that has to be overcome. The finding that usually only a few features distinguish texts by the same author points to an earlier finding by Mosteller and Wallace [2008], specifically that one should not rely on only a few features to determine questions of authorship.

However, as Koppel et al. [2011] point out the simplest type of authorship attribution problem and which incidentally has received most attention is the ideal case of having a closed set of candidate authors and copious quantities of text by each candidate as well as the anonymous test piece. Even the open-class 'unmasking' considered by Koppel et al. [2007] relies on sufficiently long samples to return accurate results, which might be an unrealistic requirement, especially in, for instance forensic contexts. Therefore, Koppel et al. [2011] consider a novel attribution method addressing what they conceive of as the three most common deterrents to using common authorship techniques [Koppel et al., 2011, p.84]:

1. "There may be thousands of known candidate authors."
2. "The author of the anonymous text might be none of the known candidates."
3. "The 'known-text' for each candidate and/or the anonymous text might be very limited."

They consider a set of blog posts, extracting 2000 words of known text and a 500 words long snippet (from the end of the post). They use a similarity-based approach (cosine similarity) on space-free character tetragrams. The task is to find the right author of a given text snippet,

based on evidence from varying feature sets, the rationale being that only the right author is going to be consistently similar to his own *unknown* piece. An author is selected only if his being the top match exceeds a particular proportion or threshold, otherwise the method returns a 'Don't know' answer. Unsurprisingly, a greater number of feature sets and a closed-candidate set yield greater accuracy, i.e. 87.9% precision at 28.2% recall. Interestingly, in the closed candidate setting, reducing the number of candidates improves accuracy (e.g. 1000 candidates yields 93.2% at 39.3% recall), while in the open-class setting having fewer candidates actually introduces issues in that an author might end up erroneously being chosen because he has less competition. Overall, the authors find that their methods achieve passable results even for snippets as short as 100 words, but note that there is still no satisfactory solution for the case of a small open candidate set and limited anonymous text.

Author Profiling Another dimension in authorship studies is to, rather than use an author's words to distinguish his or her documents try to infer information about him or her from the text, specifically aspects of inherited traits, such as gender or age, or acquired traits, such as aspects of one's personality. Being able to predict certain characteristics of an*unknown author* is not only of interest to forensic linguistics, but also in the context of, for instance discerning a companies' target customers from anonymous product reviews [Rangel and Rosso, 2013].

Predicting a certain author's characteristics, such as age or gender based on his or her text samples has been studied extensively as part of the PAN competitions (see for instance [Rangel and Rosso, 2013; Rangel et al., 2016]). While Rangel and Rosso [2013] only addressed predicting age and gender of the author, Rangel et al. [2016] considered the same task across different genre. The task of how gender and genre type interact had already been addressed previously by Koppel et al. [2002], who used a machine-learning approach on function words and part-of speech tags to study differences in performance for predicting gender in the BNC (British National Corpus), both when considering fiction and non-fiction texts on their own and conflated into one. While they are able to achieve 80% using a combination of both feature types, higher accuracy is obtained when considering each genre separately. The authors report that men's language seems to be characterised by the use of determiners, numbers and modifiers, while women's language focuses more on negation, pronouns and certain prepositions (*for* / *with*). The authors find that what distinguishes fiction and non-fiction in particular is that although men use significantly more determiners than women in both genre, in non-fiction women use the most frequent determiner *the* with about the same frequency as men.

There has also been considerable research based on counts over selected human-evaluated content and style words. LIWC (Linguistic Inquiry and Word Count) was originally conceived to better understand how people use language in the context of emotional aspects in their lives [Pennebaker et al., 2001]. It was developed by having human judges rate the degree of relatedness of particular words to both less ambiguous categories of standard function words, such as pronouns or prepositions and more vague concepts, as for instance *negative* and *positive*

emotion words. Chung and Pennebaker [2007] present their findings with respect to a very informative subset of function words: pronouns. Based on a study of 1-2 years worth of daily diaries of one biological man and one biological woman, the authors find that testosterone injections suppress the participants' use of non-I pronouns with the effect levelling off and the two participants making more references to other humans as the testosterone levels drop. The authors also note that although the use of *we* or *us* rather than *I* or *me* is culturally often conceived as reflecting the speaker's close emotional ties to others, findings on this are relatively inconclusive, with especially males using the *we* in a distancing or 'royal-we' form. According to Chung and Pennebaker [2007], the use of 1st person singular could also be linked to depression as reported by studies on depressed students contrasted with both formerly and never depressed students. Considering two people in an interaction, the one with the lower usage of 1st person singular pronouns tends to be the higher status participant [Chung and Pennebaker, 2007]. Other factors named are cultural influences beyond the self-focus and collective-focus division, such as uncertainty avoidance. As for gender differences, Newman et al. [2003][1] have found that undergraduate females tend to use 1st person singular pronouns at a consistently higher rate than undergraduate males, a finding which the authors relate to women being generally more self-focused than men, their being more prone to depression than men as well as traditionally having held lower status positions relative to men. All considered, especially the use of 1st person singular pronouns could be caused by different factors the exact relationship between which does not seem entirely clear. In addition, many of the above findings are based on very small subsets of the population, elsewhere referred to as WEIRD societies (**W**estern, **E**ducated, **I**ndustrialized, **R**ich, and **D**emocratic) unlikely to be representative of the behaviour of the general population at large [Henrich et al., 2010]. In the context of stylistic analysis, there are a few possible factors that might unnaturally inflate the usage of particular type of pronouns, such as narrative perspective (first-person vs. third person perspective), text type (dialogue vs. simple narration) or genre, and also whether the given text is likely to be autobiographical. This indicates that simply adopting a particular type of narration might not necessarily link back directly to the author's personality traits.

One inherent issue with categorising words, as for instance emotion words out of context and one which the researchers themselves acknowledge (e.g. [Chung and Pennebaker, 2007, p.345]), is that their meaning is influenced by presence of negation or use of irony or sarcasm, rendering simple word count programs essentially probabilistic. These confounding factors also present one of the more difficult challenges in the field of opinion mining and sentiment analysis. Pennebaker and Stone [2003] considered how age influences certain aspects of language; as this study pertains more strongly to individual language change, it is discussed separately in section 2.5.2. An important aspect to consider is how gender-specific features change over time and which of them prove to be reliable, since for instance Schler et al. [2006] reported the existence of an age-effect regardless of gender in the context of web blogs.

[1] This was cited by Chung and Pennebaker [2007].

Plagiarism Detection Other more subtly-related tasks with the realm of authorship analysis are those of plagiarism detection and spam filtering. One can see general similarities between the tasks of plagiarism detection, spam filtering and authorship identification (verification), as all aim at uncovering fraudulent behaviour and detection of original vs. derived work through a classification task [Oakes, 2014]. Plagiarism, more specifically can be defined as the 'unacknowledged use of another author's original work' (Martin Potthast et al. (2009) cited by [Oakes, 2014, p.59]). The type of detection applied relies to some extent on whether external sources (e.g. suspected plagiarised documents or websites) are available to the researcher. If no outside sources to compare to are accessible, the suspicion of a text having been plagiarised has to be investigated intrinsically, by considering inconsistencies within, i.e. by comparing individual passages to the entire text, a task complicated by small document sizes. It thus bears similarities to methods in disputed authorship in that stylistic features have to be applied [Oakes, 2014]. In contrast, methods in extrinsic plagiarism detection encompasses a wider range of possible options targeted at different types of plagiarism or text reuse. The type of plagiarism easiest to detect is use of an exact copy of the source document, as the suspicious passage might only have to be subjected to a search engine. In many cases, however, the original plagiarised text has been obfuscated, e.g. has been rewritten or paraphrased and reordered [Oakes, 2014]. This obviously renders the task of detection more difficult and requires the use of more intricate methods, as proposed by Chong and Specia [2011], who used Wordnet synset overlap by computing the number of common 5-gram synsets in two documents. As with all classification tasks, one has to decide what sequence size to consider, from the bag-of-words approach commonly employed in search engines to longer more specific ngrams not usually found in all texts. One very central aspect to assessing similarity between texts to detect plagiarism is to determine different levels of similarity, as was considered by Hoad and Zobel [2003], who reason that no document should be rated closer to a suspicious document than it would be to itself and find that independent texts written on similar topics rate very low in comparison. This is reminiscent of the lining up of *impostors* technique introduced by Koppel et al. [2007]. Generally, plagiarism detection methods may target individual sub-parts of the entire document which is probably most suited for cases where it is believed that only isolated parts are plagiarised supplying a score to each part separately. Otherwise the similarity of one document to another could be expressed as a global measure of similarity as was chosen by Hoad and Zobel [2003].

2.2.2 Document Dating

Another type of text classification is 'document dating'. This refers to the manual or automatic analysis of texts with the aim of determining when it has been written. The following discussion focuses on research with the sole purpose of *dating a document* rather than for stylometric or language change purposes and I defer treatment of those works to section 2.7. There is, how-

ever, considerable overlap between methods for document dating and discovery of diachronic features but for the general motivation, and hence other natural divisions might be more suitable given other research contexts. Similar to a literary scholar trying to estimate an author's characteristic style, document dating can be undertaken manually by a literary expert either using internal text evidence, i.e. the usage of particular words or phrases or external evidence in the form of witness accounts or properties of the work itself, e.g. the paper or ink used. One example of a manual document dating is 'The Donation of Constantine', allegedly a decree of Roman Emperor Constantine I (272–373), that was declared to be a forgery by Catholic priest Lorenzo Valla, who deemed the language to not have originated in the 4th century but rather the 8th century. Frontini et al. [2008] revisited the question of *The Donation of Constantine*'s temporal origin, hereafter 'DOC', by considering homogeneity of Latin texts over seven temporal periods, from early Latin before 100 B.C to modern and contemporary Latin. The authors assess the DOC's homogeneity fit with its alleged origin (period 3) and its actual origin (period 4) based on letter bigram sequences. Based on their findings, they conclude that as internal analyses are compatible with the possibility that the document is not a forgery, but external evidence suggests that it was in fact one, it must be a very good forgery in terms of morphological similarities as these are usually outside the author's conscious control. Another work on historical text dating focused on three different historical corpora; the Corpus of late modern English texts (CLMET), the Portuguese Colonia corpus and and a Romanian diachronic corpus [Niculae et al., 2014]. The dating of documents is done by a ranking approach using ordinal regression based on occurrence counts of lexical features and character ngrams in final positions of words. The authors report high pairwise accuracy rankings of 83%-93%. While the previous approaches to document dating have exploited stylometric features for estimating a text's temporal origin, there has also been research based on the other side of the frequency spectrum, i.e. rooted in information retrieval. Kotsakos et al. [2014], for instance, exploit sudden frequency bursts of characteristic terms linked to specific events, such as *earthquake* or *shooting* to determine timestamps of texts accurately, reporting consistently higher precision and speed of performance than the state-of-the-art methods.

2.2.3 Stylistic Homogeneity

One of the studies mentioned in the previous section, Frontini et al. [2008], considered the task of dating a document by examining homogeneity conditions in diachronic corpora. The property of homogeneity within corpora used for comparing word distributions lies at the core for making valid assumptions on their basis [Kilgarriff, 2001]. Yet, it has been shown that comparing two distributions based on randomly-allocated words from the same corpus can result in significant differences [Kilgarriff, 2001]. Comparing corpus properties to those of randomly-generated distributions is only marginally helpful, as language has been shown to be non-random, causing issues especially for estimating relations between very rare linguistic

events [Kilgarriff, 2005]. Difference between two subsets of the same corpus might also be reflected in the very frequent words, rather than only in the less reliable *hapax legomena* or *dislegomena*, rendering the former not impervious to varying distributions.

One manner in which a corpus could be heterogeneous (or *not homogeneous*) disregarding differences based on genre is through temporal influences. Regular or synchronic analyses of style usually simply assume that temporal influences are random and that lack of homogeneity would inherently be due to random fluctuations. While the very deterrent of language change presents an issue for synchronic analyses, lack of homogeneity is less problematic for diachronic analyses that are built on the assumption that from one time instance to the next, there might not be homogeneity. However, temporal analyses could fall prey to over-estimating heterogeneity or differences between two time instances. This effect is lessened by taking into account the entire time span, where non-random change in the form of, for instance an upwards trend gives higher validity to individual observations. Another remedy to focusing too much on the very changeable aspects usually represented by very high frequencies in isolated texts, is to disregard items that do not occur in all time instances of a time line considered, thus emphasising typical distributions and more representative aspects of authorship in the data. Although these measures would not invariably protect against flawed inference, they might lessen the effect. Nevertheless, any analysis has to content itself with the fact that "[B]etween any two bodies of text, some of the enormously many measurable variables will consistently be similar solely by chance, even if both texts were written by different people; conversely, some features will be consistently different, even though the works have common authorship."(David Banks cited by [Rudman, 2003, p.28]).

2.3 Feature Analysis

The following section considers properties of features commonly used in studies of authorship. Section 2.3.1 gives an overview of the types of features commonly used in this domain and discusses further aspects and implications of this. Section 2.3.2 then continues by addressing aspects of meaning.

2.3.1 Authorship Features

The features traditionally used for stylometry or authorship attribution differ fundamentally from those employed in other areas of text classification, such as information retrieval. Most classification tasks are focused on grouping texts according to some characteristic that relies strongly on the content words within, e.g. deciding what documents are closest in topic to a query term in an information retrieval setting. In this context, words common to all documents examined would not be helpful in discriminating between them and are often discarded completely. In contrast, style analysis in authorship often only considers the very frequent

(grammatical) words in a text because these are common to all texts and one of the few means of comparing texts with potentially very different topics. Whereas in information retrieval, one would like to characterise a text to identify the words both frequent within but ideally not occurring (much) in any other texts, in style analysis the difference between texts of two authors lies primarily in differing rates of the same words or features in general. Historically speaking, Mosteller and Wallace [2008] were among the first to use and popularise function words for authorship attribution, as they appeared in all texts regardless of topic and genre and as they reasoned were not subject to conscious control by the author. One has to make a distinction between the features that are frequent because they are grammatical and document-based features, such as average word or sentence length, that can (almost) always be used regardless of the language and genre examined. Very early attempts at the authorship attribution task made use of one of these more general document-based features in the form of word length histograms [Mendenhall, 1887, 1901] (see 2.1). Since then both global measures of sentence and word length have not been found to discriminate well in authorship attribution tasks [Oakes, 2014, p.9] and have been argued to be better suited for discriminating between register or genre [Oakes, 2014, p.5]. Other measures usually based on the entire document are measures pertaining to lexical richness, such as type-token ratio (TTR)[2] and proportion of *hapax legomena*.

However, the number of different 'types' that are also often referred to as 'vocabulary size', depends heavily on text length where as "the text length increases the vocabulary also increases quickly at the beginning and then more and more slowly" [Stamatatos, 2009, p.5]. Tweedie and Baayen [1998] consider the question of measures of lexical richness in a more comprehensive and systematic fashion by evaluating the reliability of a given measure with respect to two basic aspects that are subject to text length: "Firstly, is a given statistic mathematically constant, given the simplifying, but technically convenient assumption that words are used randomly and independently?" And "[s]econdly, how is a constant affected by violations of the randomness assumption in actual texts?" [Tweedie and Baayen, 1998, p.324]. They consider a plethora of different measures of lexical richness, which they test on both a theoretical distribution based on the assumption that all words occur independently and an actual empirical distribution, finding that most measures that are theoretically constant given the word's independency condition are not so given an actual distribution, e.g. a text exhibiting different stages of vocabulary growth. They also test the measures with respect to stylistic analyses for distinguishing between different authors, finding analyses based on function words to be more accurate. Further, they identify two major families of vocabulary measures; the ones capturing lexical richness and the ones capturing repetition rate. They conclude that the assumption that measures of lexical richness are independent or even roughly independent of text length is invalid given that all their tested measures changed substantially and in systematic ways with text length, making it necessary to correct for text length or consider developmental profiles of the whole text. One

[2]One needs to differentiate between type and token frequency, where types are unique word forms or constructions and tokens refer to their actual frequency in the text.

obvious remedy for this would be to only compare texts based on the same text length or only consider measures that account for differences in length, yet according to [Oakes, 2014, p.6] in the context of authorship analyses this would still not suffice as type-token ratios vary a lot between authors and even within their texts.

Another distinction with respect to feature types that has to be made is whether a feature type is 'linguistically-informed', i.e. has some linguistic-psychological grounding rather than being 'non-linguistically-informed'. This aspect has to be differentiated from psychological control on the basis of the author: features can be outside an author's conscious control while still being linguistically-motivated, e.g. higher rates of self-referring pronouns is argued to be indicative of depression [Pennebaker, 2011]. Non-linguistically motivated features would not exhibit this type of characteristic, such as character unigrams that are sometimes considered an *intelligent* baseline. Somewhat surprisingly and despite the lack of intuition about them, character ngrams have been found useful and accurate especially in language-independent tasks including authorship attribution. Unlike lexical features, character ngrams are also more tolerant to noise, such as grammatical errors or unusual punctuation [Stamatatos, 2009]. Character ngrams usually include punctuation and spaces between words. As the ngram size grows to $n > 4$ approaching average word length, the features conceivably become more meaningful and linguistically-motivated. Stamatatos [2006] reports very good results in using character ngrams of sizes 3-5 in an author identification task based on ensemble-learning.

However, the probably most widely used features in authorship analysis are the most frequent word unigrams as used extensively by Mosteller and Wallace [2008]. Usually these features are applied in a bag-of-words approach reducing the text to simple word distributions not retaining any contextual information. This approach may introduce ambiguities into the analysis, as the same lexical representation may appear in different syntactic contexts and depending on this would also be allocated to different word classes. This aspect of one lexical representation (or 'signifier') having the potential to refer to two different referents ('signified') may primarily be an issue for languages with less inflection, such as English, that causes there to be more overlap between items. For instance, the word *like* can appear in the position of verb, noun, preposition, adverb and conjunction and although most of its uses would be somewhat related, these might still have different semantics depending on context, e.g. *to be* $\underline{like_{IN}}$ *someone* is comparable to saying *to be* $\underline{similar}$ *to someone*, whereas *to* $\underline{like_{VB}}$ *someone* is closer to *to be* $\underline{fond}$ *of someone*.[3] This distinction has been made in some stylistic analyses, such as by Burrows [1989] who differentiates between different meaning in contexts of the items *to*, *that*, and *for*. Thus, even though this is dealing with syntax rather than semantics, the perception of an author's style could be somewhat different if not making this distinction apart from the more general issue of not capturing word class usages accurately. While the proposed word unigram representation might be more accurate, using syntactic information for meaning enhancement makes one strongly reliant on the accuracy of one's parser [Stamatatos, 2009]. This

[3]Throughout this work, the Penn Treebank tagset is used to refer to specific part-of-speech tags.

is equally true for only using syntactic information in the form of part-of-speech(POS) tags or output from a stemming method that reduces each word to its stem. Both part-of-speech tags and word stems offer abstractions away from the complete word, where POS tags focus on the syntactic attributes of texts, and word stems might capture slightly more of the words' semantics.

However, syntactic features are arguably even more subconscious than function words as they are not lexicalised and therefore represent highly automatic and unconscious attributes of language production [Chaski, 2001].[4] Zhao and Zobel [2007] tested POS unigram and bigrams in an authorship attribution task on a multiple author corpus, but found function words to be better discriminators, which they partly attribute to tagging errors. Another type of syntactic feature are syntactic rewrite rules, as for instance used by Baayen et al. [1996] in an authorship attribution task to tell apart the works of two crime fiction authors Allingham and Innes.[5] Throughout this work, they compared their results to a baseline classification with a more traditional setup using the highest word frequency strata (both with annotation for syntactic function and without). For methods of comparison, they considered five measures of vocabulary richness: Yule's *K*, Simpson's *D*, Honoré's *R*, Sichel's *S* and Brunet's *W*, where *K* and *D* target the higher frequency items, and *S* and *R* depend on lower frequency items, such as *hapax legomena* and *dislegomena*, and *W* measures vocabulary richness. For the main experiment, the five measures are computed based on syntactic rewrite rules, where Baayen et al. find that the 50-most frequent rewrite rules satisfy both their criteria for accurate authorship attribution: correct assignment of all test samples as well as good separation of all known samples. Further, they also investigate the accuracy of rewrite rules at the other side of the frequency spectrum, the *hapax legomena*. For this they use a measure of 'the rate at which new items appear' thereby addressing syntactic creativity. In order to increase author-specific sensitivity, they select those left-hand sides with most different right-hand side realisations. They find that this technique proves to be more robust than the techniques focused on the highest frequency rewrite rules. Considering development profiles, they also find that the use of syntax is more uniform than that of word usage. Narayanan et al. [2012] explore various feature types in a large-scale authorship identification task and find that while syntactic pairs based on parse trees' parent and child nodes, perform equally well as character unigrams or function words, while able to boost performance in certain settings.

To reiterate, when selecting style markers, especially in cases of disputed authorship, it is essential not to cherry-pick specific markers as they happen to separate texts well, as any two texts either by the same or different authors will have both features that are very similar and that are very different between them (David Banks cited by Rudman [2003]).

[4] This research is limited to part-of-speech features, as the main focus is on developing methods for analysis rather than providing an exhaustive investigation into adequate features in this context.

[5] A preliminary analysis as part of this study by Baayen et al. [1996] had shown that register poses as a serious confounding factor where authors can be more similar to other authors within the same register than to themselves in a different register.

2.3.2 The Meaning of Features

Words are often allocated to either content or function type, thus being broadly classified as either topic-dependent or topic-independent, although a strict dichotomy is often neither practical nor representative, seeing that members of both function and content categories share characteristics with the respective other category. For instance, some very frequent verbs, such as *to be* enjoy dual status of both auxiliary and main verb and likely due to their high token frequency they retained their irregular inflection over time [Lieberman et al., 2007]. Other more abstract examples include frequent temporal expressions consisting either only of nouns or of adjective and noun combinations, such as *(last) year*, *(last) month* or *(next) summer*. One could argue that these, although belonging to a (very productive) content category might in fact be closer in frequency and behaviour to the function categories. For this reason, it might be more authentic to consider words based on their frequency behaviour and occurrence patterns, suggesting a more fluent, continual representation rather than a strict categorical one often implicated in most research. This suggests classification of features along two different dimensions: the 'relative frequency' dimension and the 'occurrence' dimension. Features can be rated according to their (mean) relative frequency and according to the proportion of individual texts out of an entire text collection they would typically appear in. This gives rise to four possible combinations simplifying the more comprehensive continuum-based representation to a categorical representation with categories: 'frequent', 'infrequent', 'occurring in all text', 'occurring in few texts'. While at opposite frequency spectra, there are function types (mostly frequent and occurring in all texts) and *hapax legomena* (highly infrequent and only occurring in one text), one can also imagine the other combinations, i.e. words infrequent but occurring in all texts, such as more unusual function words as well as words that are very topic-dependent and thus frequent in only a few texts (frequent and occurring in few texts). Similarity on both axes likely renders features more similar to each other regardless of whether they share the same word class, as it signals similar behaviour in texts.

The Meaning of Co-occurrence The implications of word co-occurrence can be manifold ranging from relatively *meaningless* to being reasonably *meaningful*. As already discussed earlier, in general word distributions are non-random [Kilgarriff, 2005], and both grammatical and content properties influence their realisations. Overall, one has to broadly differentiate between two different types of co-occurrence: the one that is measurable through actual *proximity* in the text, e.g. two items are part of the same noun phrase and the one where two words may just occur in the same text, making reasoning about their possible relatedness harder to empirically measure. The most common type of co-occurrence and incidentally the one most easily measurable is a 'collocation'.

A collocation typically consists of two or more words that correspond to a common way of expressing something. Collocations are special in that they are characterised by limited com-

positionality, i.e. very often the expression's entire meaning cannot be completely predicted by the meaning of its parts, such as *wide awake*, which would be mildly compositional to *let the cat out of the bag*, an example of an idiom, which as a group tend to be non-compositional [Manning and Schütze, 1999]. Usually, every part of the collocation is fixed, so that no word can be substituted by a synonym and achieve the same meaning. Words that form a collocation need not occur right next to each other in the text, for instance, phrasal verbs, such as *put up* could also be separated by a noun phrase argument: *She put the guests up for a week*. Some definitions classify idioms, phrasal verbs and collocations as 'multiword expressions' (MWE) and reserve the term 'collocation' for adjective-noun or noun-noun sequences only. Manning and Schütze [1999] find in their experiments that even high frequency and low variance of two words does not clearly indicate a collocation, as there is a lot of chance co-occurrence. Lyse and Andersen [2012] consider different statistical association measures to detect multiword expressions in Norwegian newspapers by ranking bigram and trigrams for their tendency to co-occur. In this, they differentiate between various types of MWE, e.g. foreign MWE, grammatical MWE or idiomatic phrase, finding that association measures tend to only return good results with a subset of the groups. For instance, both log-likelihood and t-score seem to favour grammatical and thus more frequent MWEs. They report that none of the measures considered is particularly good at spotting idiomatic expressions. Jurafsky et al. [2001] examine how words on either side of a target word can influence phonetic reduction. They investigate these probabilistic relations between function words and content words based on the spoken *Switchboard* database. With respect to function words, they find clear influences onto the target word by both the word preceding it and the word following it in that if those are more predictable the target word is more likely to show vowel reduction and shortened speech duration. The investigated effects of content words' *final-t/d* deletion showed that content words with higher relative frequencies are shorter and more likely to have deleted *t* or *d* than words with lower relative frequency, but content words' results for conditional probability were much weaker than those for function words. This research, although on spoken, rather than on written data is relevant in that it considers influences of neighbouring words quantified by probability relations. Words that occur close to each other in texts can be more easily examined for associations between them, whereas the task becomes considerably more difficult if one cannot reason based on actual proximity, but possibly only based on occurrence in the same paragraph or text. One possible way of measuring relatedness of words over time would be to analyse change in relative frequency patterns that if similar could conceivably hint at words being more related.

2.4 General Language Shift

When considering aspects of language change it is important to make a distinction between 'function' and 'form'. For instance, studies trying to detect neologisms focus on a change in

function of a particular word form. This is obviously more difficult to detect than the simple emergence of a new word form, as one has to decide whether a particular lexical representation has acquired a new function based on possibly new contexts it occurs in, different frequency distributions over time or external evidence. This research concentrates mainly on analysing the form, thus it pays less attention to items acquiring a new sense, where this would not be visible through a change in word class.

One important distinction in this is the differentiation of type and token frequency, where types are unique word forms or constructions and tokens refer to their actual frequency or occurrence in the text. Differences in token frequency do affect whether a particular word form is more likely to be subject to 'regularisation' as studied by Lieberman et al. [2007] in the case of irregular verbs acquiring the '-ed' past over time. The authors find that the more infrequent an irregular verb is, the faster it will regulate to the *-ed* past tense form. However, very frequent verbs, such as *have* or *be* are less likely to be regularised due to their high entrenchment.

Other areas of linguistic research have also considered the change of broader categories of words, such as frequency effects in syntax, largely distinguishing between type and token frequency of particular variables or categories [Bybee and Thompson, 1997]. Bybee and Thompson [1997] discuss three frequency effects that are important not only in shaping phonology and morphology, but also syntax; two effects are caused by *high token* frequency, which have adverse tendencies that can only be explained by considering the influence of the third frequency effect of high type frequency. A high token frequency of an item promotes its 'reduction', as visible in conventionalised contractions in English (*I'm*, *can't*). In contrast, the 'Conserving Effect' is visible with high token items, where the more the form is used the more it is strengthened, compare normalisation of the English past tense of *weep* from *wept* to *weeped*, compared to high frequency items, such as *sleep* (*slept*). A syntactic example of this is the fact that pronouns, although derived from full noun phrases, show much more conservative behaviour (e.g. case marking) due to their higher frequency.[6] The type of change that is resisted in the high token frequency items is change on the basis of combinatorial patterns or constructions that are *productive*. "The more lexical items that are heard in a certain position in a construction, the less likely it is that the construction will be associated with a particular lexical item" [Bybee and Thompson, 1997, p.384]. This is observable in the ditransitive construction, which is only acceptable with very specific lexical verbs of high frequency, compare: *He told the woman the news* vs. **He whispered the woman the news* [Bybee and Thompson, 1997, p.385]. To a limited extent, this is also productive, in that the construction can apply to a few new verbs, such as *e-mailed* or *telephoned*. Bybee [2008] consider grammaticising items, such as *can*, which evolved from having more specific meanings to becoming semantically 'bleached', making them appropriate for usage in several different contexts causing its co-occurring lexical items to increase in type frequency, whereas the construction itself increases in token frequency

[6]The more conservative behaviour of pronouns and by extension other function words renders them good, reliable candidates for temporal analyses.

[Bybee, 2008]. An example of this is the construction *going* + infinitive that can now be used to denote future tense, but used to be a mere movement verb, such as *travelling* or *riding*. Bybee [2008] attributes this effect of grammaticalisation of certain constructions to its higher frequency or rate of repetition, which supposedly weakens the semantic force, creating greater autonomy and thus making it more entrenched in the language.

Hamilton et al. [2016] consider the function aspect of diachronic change by taking a closer look at *global* and *local shifts* in a word's distributional semantics. Based on previous findings in the field, they hypothesise and find that local or cultural shifts' are more pronounced in nouns, whereas verbs more readily participate in global or more regular processes of semantic change. For the global change, they compute the cosine distance between two word vectors capturing the co-occurrence statistics at consecutive time points t and $t+1$. For the local measure, the word set is limited to a word's n nearest neighbours (according to cosine-similarity). They use the similarity values from both the local and global setting as response variables in a linear regression context taking word frequency, the decade of change and a variable indicating noun or verb type as independent variables. The data sets comprised both Google ngram sets (1800–1990) that have large amounts of historical text from English, French and German and the Corpus of Historical American English (COHA) for the years 1850–2000. They report the differences between local and global measurements for three well-attested regular linguistic shift words (i.e. *actually*, *must*, *promise*) that changed more according to the global measure and three well-known examples of cultural changes (i.e. *gay*, *virus*, *cell*) that as the authors report change more according to the local neighbourhood measure. Across all languages, as predicted, the local neighbourhood measure assigns higher rates of semantic change to nouns than verbs with the opposite being true for the global measure. This also remains the case, when adverbs and adjectives are included among the verbs, supporting previous results in the literature suggesting that adverbial and adjectival modifiers are often the target of regular linguistic change [Hamilton et al., 2016, p.4]. Although the research described in this thesis largely disregards the semantic change aspects, it is an important confounding factor that should be controlled for, even when it is not the focus of the investigation. Thus, finding a variable drop or rise significantly from one year to the other could be more indicative of semantic change rather than only sudden popularity.

The research described by Štajner and Mitkov [2011] investigates diachronic changes in 20th century written British English (BE) and American English (AE).[7] They consider four different variables: average sentence length (ASL), Automated Readability Index (ARI), Lexical density (LD) and Lexical Richness (LR)[8] across four different text categories: *press*, *general prose*, *learned* and *fiction*. The changes between the years were assessed using a two-tailed t-test, taking the 1961 sample as starting year while also measuring the percentage of change

[7] Published: 1961(BE+AE) and 1991(BE)/1992(AE)

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

based on this value in each case.[9] Considering the results for British English, while the ARI shows a statistically significant increase over the observed period in the press and prose categories (interpreted by the authors as a tendency to render texts more difficult to read in these categories), the ASL did not change significantly in the period 1961–1991 in any of the four categories. With respect to lexical richness or density, it is reported that both variables increase across press, prose and fiction, this being most pronounced for the press category. As for the American English corpora, while there is no significant change in ARI, ASL decreases significantly for the press and learned text categories, which is interpreted as an example of *colloquialisation*. Lexical richness and density only increased in the prose text category. Comparing British and American corpora for both start and end year for each genre and feature separately reportedly support the previous hypotheses. The press category shows some interesting results in that while the difference between AE and BE are still significantly different over ARI/LD/LR in the 1961 time slice, this significant difference has disappeared in the 1991/1992 comparison, which is attributed to the growing *Americanisation* that would be particularly tangible in this category. In contrast, ASL actually became more significant as AE decreased for this over time while BE did not.

While this study considered broader linguistic categories presumably more stable than, for instance individual words, there remains the question of what changes possibly occurred between the years examined and which would indicate whether the current state and trend is likely to be permanent. Thus, considering individual time slices cannot take into account changes that occurred in between that period, especially when the frequency draws nearer to its original or starting value.

2.5 Language Decline and Linguistic Ageing Effects

In addition to being subject to general language shifts or background language change, to some extent an author's style may also subject to factors of linguistic ageing or even premature language decline as for instance caused by dementia or Alzheimer's disease. Section 2.5.1 considers research analysing pathological decline in language, while section 2.5.2 then addresses studies dealing with regular ageing effects in language.

2.5.1 Early Language Decline

In order to investigate effects of dementia on individuals' ability to produce written or spoken language, one needs longitudinal data ideally following individuals' linguistic development over a few decades. One of these longitudinal studies is the *Nun study* [Kemper et al., 2001;

[9] The t-value resulting from a comparison between two means measures the size of the difference between an observed sample statistic and its hypothesised 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 hypothesised and observed value.

Snowdon, 2003]. This study focuses on ageing and Alzheimer's disease starting in 1990-1993 with all 678 participants being members of the School Sisters of Notre Dame congregation. The language samples for each participant are based on annual assessment after joining the study and original, handwritten autobiographies written at the time the participants took their vows, thereby providing an early individual language sample. In addition to the linguistic evaluations, annual examinations for cognitive and physical function pertaining to dementia factors were carried out for each patient. The advantage of this particular design is that other than regular observational studies that contrast patients with healthy controls trying to discern what particular aspect of, for instance lifestyle or genetics was responsible for causing the disease, it is less likely to focus on factors that just happen to separate healthy and sick participants, especially since all the study's participants are more comparable with respect to living factors. This longitudinal data collection gave rise to a plethora of research articles of which only a select few are going to be discussed in the following.

The research described by Kemper et al. [2001] contrasts participants who were to develop dementia against those who were not. The principal dimensions considered here were 'idea density' and 'grammatical complexity'. Idea density can be approximated by considering the ratio of open-class categories of words to the total number of lexical tokens. "[L]ow idea density in young adulthood may reflect suboptimal neurocognitive development, which, in turn may increase susceptibility to age-related decline due to Alzheimer's or other diseases"[Kemper et al., 2001, p.227]. Grammatical complexity is often measured by considering the proportion of complex clauses (containing dependent clauses) to the total number of clauses. While idea density seems to be correlated with measures of vocabulary, grammatical complexity is correlated with measures of working memory, including digit span and reading span [Cheung and Kemper, 1992].

The authors use a linear mixed model to investigate both differential change effects (between-subject) and patterns of change (within-subject) with respect to idea density and grammatical complexity. The authors found that both variables declined over time for both groups although at different rates: participants who were to develop dementia had a lower start rate and a lower rater of decline for those two variables observed. For participants who did not meet the criteria for dementia, grammatical complexity declined 0.04 units per year from an average of 4.78 to 2.34 (on a 0-to-7-point scale). Idea density scores decreased 0.03 units per year from an average of 5.35 propositions per 10 words to 3.52 propositions per 10 words. The decline in dementia participants happens at a slightly slower rate, 0.03 from 3.86 and 0.02 from 4.34 for grammatical complexity and idea density respectively. Therefore, the main difference between participants not meeting the criteria for dementia and the ones who did, was the initial level: 4.78 vs. 3.86 for sentence complexity and 5.35 vs. 4.34 for idea density. The authors note that their analysis was not sensitive to any departures from linearity as only two intermediate time periods were available leaving over 30 years uncovered.

While the research described by Kemper et al. [2001] dealt with autobiographical accounts

in part written specifically for the subsequent linguistic analysis, Le et al. [2011] aim to detect markers of dementia in the writings of three female British novelists. In particular, they contrast Iris Murdoch, who died with Alzheimer's disease, with Agatha Christie, who was suspected of having it and P.D. James, who has aged healthily. The authors consider a variety of lexical and syntactic measures based on earlier research suggesting that vocabulary and syntactic complexity declines more rapidly in dementia, especially the use of low-frequency and more specific words, as well as lexical repetitions and disfluencies. Also, the passive voice is supposed to be an indicator of faster linguistic decline, in that fewer passive constructions were used by the non-healthy group, as well as simpler agentless passives. Several works across each author's creative lifespan were analysed for different indicators of vocabulary and syntactic complexity strength over time. The hypotheses with regard to more rapid lexical decline in Murdoch are largely confirmed. 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 in tact throughout. Christie's lexical types all show an overall decline with only two exceptions. The vocabulary, repetition and specificity scores vary only on a very small scale in James' novels. Thus, the authors note that although Murdoch does not share Christie's increase in indefinite nouns, they both show common lexical decline not found in James validating their hypotheses with respect to lexical markers. The results of the syntactic complexity analysis is somewhat puzzling: while no significant linear trends can be found for Murdoch over the entire period, there occurs a drop in her 40s and 50s, followed by a less intuitive period of recovery and then a slight decline for her last two novels. Christie's results vary to a great degree overall indicating a rising rather than declining tendency. Both Murdoch and Christie show an overall decrease in passive constructions, but a proportional rise in simpler (*get-passive*) constructions and a drop of the more difficult (*be-passive*), although not all results are significant. However, Murdoch's agent-passives increase significantly. James' syntactic results vary only slightly.

Although the analysis and data preparation was very carefully conducted, as the authors note, the data set is somewhat small with only 1-2 people for each of the two conditions, and it is therefore unclear what aspect of the results are completely reliable, as each of the examined authors could potentially be unrepresentative of their group. In addition, ideally the general language shift and possible stylistic change should also be taken into account to exclude possible confounding factors. Some previous results in the literature were confirmed, while others could not be replicated, showing that more research in the area is needed.

One of the issues facing researchers in either area of (temporal) stylometry, linguistic ageing, pathological linguistic decline (and general language change) is how to disentangle one's research interest from the other three dimensions and how to be certain that what is left is really due to the effect one was investigating.

2.5.2 Linguistic Ageing Effects

While the studies considered in the previous section considered symptoms of pathological linguistic decline, the study by Pennebaker and Stone [2003] focused on aspects of regular and to be expected linguistic ageing. In particular, they posed four hypotheses about the effect of ageing on language, [Pennebaker and Stone, 2003, p.294]: the first one hypothesised that ageing should be a associated with drops in negative affect words and slight increases in positive affect words. The second hypothesis speculated on a decrease in the use of social words and first-person plural pronouns to the degree that individuals reduced their social networks with ageing. The third hypothesis focused on the usage of tenses, specifically 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 would become apparent with increasing age. The fourth hypothesis dealt with cognitive complexity and to what extent it would extend to language, i.e. older ages were predicted to use fewer cognitively complex words. Specifically, they expected a quadratic relationship between cognitive markers (cognitive mechanisms and causal, insight, and exclusive words) and age, whereas markers of verbal ability, including the use of large words, were not expected to show monotonic increases or decreases over time.

The two data sets considered for this were a corpus containing self-reports from psychological disclosure studies and a corpus with collected works of 10 authors over time (between the years of 1591–1939). The Disclosure study 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]). The statistical analysis for this included correlation analysis and both simple linear and quadratic regression.

The second set contained text samples of 10 different authors, both British and American, male and female, across different genre: dating was based on when a work was written, averaging over years where necessary and reverting to the publication date in cases where the composition date could not be determined. Part of the analysis was correlational, analysing the simple relationship between language use and age, the nature of the data supporting within-author correlations rather than only between-subjects' correlations as in the Disclosure project. In order to arrive at the final score shown, the Linguistic Inquiry and Word count (LIWC) scores for each author for each work were correlated with the age of the author in the year the work was written. Next, the means of the within-author correlations for each of the variables were computed and subjected to single-sample t-tests to evaluate whether each mean was significantly different from zero. Five of the original 14 correlation means were significantly different from zero ($p \leq .05, df = 9$). The other part of the analysis involved creating an ageing coefficient based on the findings of all 14 variables in the Disclosure study. This revealed that 6 out of the 10 authors examined exhibited the same pattern of language use with respect to ageing that was found in the Disclosure project (see Pennebaker and Stone [2003] for details).

As for the results, the hypothesised increase in positive emotion word and decrease in negative emotion words were significant for the Disclosure project, but not for the Author study. As concerns the second hypothesis about a decrease in first-person plural pronouns, instead a significant drop over time in first-person singular was found in both corpora (and a non-significant decrease in first-person plural pronouns existed in the Disclosure data). Rather than the hypothesised future to past tense shifts, a decrease in past-tense and an increase in future tense verbs was found in the Disclosure study with an increase in future tense also being significant for the author project. Although no change in long-letter sequences was predicted, they significantly increased over time for the Disclosure study, with this effect being present but not significant for the writers. Possible confounding factors for this study could have been general language change as discussed in section 2.4 and also individual stylistic differences and developments irrespective of any particular ageing process. The Author project combined works from ten English and American authors, that lived between the 16^{th} to 20^{th} century over various genre: novels, stories, plays and poetry. Even though these were analysed separately, given in particular the differences between genre with respect to function and content word shifts, one might doubt the comparability of these works and the conclusions based on them, especially considering that these might have been subject to different background language changes.

The research analysed so far has highlighted two main influences that could present distortions when analysing an author's stylistics: the general language shift he or she would be subject to and ageing effects that affect a person's cognitive abilities as they age, this possibly happening prematurely in the presence of dementia. The last two sections 2.6 and 2.7 deal with the last piece of the three parts that form an individual's style over time, i.e. diachronic stylistic change, by first introducing common methods in the field and then discussing previous work.

2.6 Temporal Regression Analysis

The analysis of data over time probably has its most prominent usage in quantitative forecasting analysis, which involves the (quantitative) analysis of how a particular variable (or variables) may change over time and how that information can be used to predict its (or their) future behaviour, thus inherently assuming that some aspects of the past continue in the future, known as the 'continuity assumption' [Makridakis et al., 2008]. Thus, the future value of a variable y is predicted using a function over some other variable values. These other variable values could be composed in two different ways, pertaining either to the use of a 'time-series' model or an 'explanatory' model. When considering a time-series model, the assumption is that one can predict the future value of the variable y by looking at the values it took at previous points in time and the possible patterns this would show over time. In contrast, for prediction, explanatory models focus less on interpreting previous values of the same variable, and more on the relationship with other variables at the same point in time. Consequently, the prediction of a variable y, using explanatory models, is based on a function over a set of distinct variables:

$x_1, x_2, \ldots, x_{p-1}, x_p = X$, with $y \notin X$, at the same time point $t : \{t \in 1, \ldots, n\}$, and some error term: $y_t = f(x_{1t}, \ldots, x_{2t}, \ldots, x_{p-1t}, \ldots, x_{pt}, error)$.

$$\hat{y}_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_p x_{pt} \tag{2.2}$$

The general model for this is shown in Equation (2.2), predicting variable y, where $\hat{y}_t$ refers to the estimate of that variable at a particular time instance $t : \{t \in 1, \ldots, n\}$, β_0 refers to the intercept, and β_p to the pth coefficient of the pth predictor x_{pt}. In the present case, the year of publication is usually set as the response variable, so that a model based on syntactic unigrams (relative frequencies) for the year 1880 could be defined in the following way: $\hat{y}_{1880} = \beta_0 + \beta_1(NN_{1880}) + \beta_2(NP_{1880}) + \beta_3(IN_{1880})$.

When analysing 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 Eq. (2.3), where y_{tj} is the response variable for author j at time t, x_{tj} is individual-specific random effect and A_j is the author-specific random effect; ε_{tj} represents the error term. Similarly, Eq. (2.4) shows the the same for the quadratic model, adding predictor $\beta_2 x_{tj}^2$.

$$y_{tj} = \beta_0 + \beta_1 x_{1tj} + \beta_2 x_{2tj} + \ldots + A_j + \varepsilon_{tj} \tag{2.3}$$

$$y_{tj} = \beta_0 + \beta_1 x_{1tj} + \beta_2 T_{2tj}^2 + \ldots + A_j + \varepsilon_{tj} \tag{2.4}$$

Regression models are customarily evaluated using the 'residual sum of squares' (RSS): given predicted values $\hat{y}_i$ computed by the model and observed values y_i, the RSS measures the difference between them. The smaller the RSS, the greater the amount of variation of y values around their mean that is explained by the model. This is known as the 'ordinary least squares' (OLS) fit, a model selection criterion that also forms the basis of evaluation measures, such as the 'root-mean-square error' (RMSE).

Rather than applying models based only on least squares regression, one can also use so-called 'shrinkage' models that offer an extension to regular OLS models by additionally penalising coefficient magnitudes, thereby aiming to keep the model from overfitting the data. Specifically, I use the 'elastic net', which is a combination of the two most common types of shrinkage, 'lasso' and 'ridge' regression [Zou and Hastie, 2005]. The elastic net penalises both the L_1 and L_2 norms,[10] causing some coefficients to be shrunk (ridge) and some to be set to zero (lasso), with the exact weighting between the two also being subject to tuning. In addition, the elastic net tends to select groups of correlated predictors rather than discarding all but one from a group of related predictors, as is common when using only the lasso technique. The entire cost function is shown in Eq. (2.5). As with lasso and ridge regression, $\lambda \geq 0$ controls finding a compromise between fitting the data and keeping coefficient values as small as possible, while

[10] $\|\beta\|_1$: $\sum_i |\beta_i|$ and $\|\beta\|_2^2$: $\sum_i \beta_i^2$

the elastic net parameter α determines the mix of the two penalties, i.e. how many features are merely shrunk as opposed to being completely removed.

$$\max_{\{\beta_{0,k},\beta_k\in\mathbf{R}^p\}_1^K}\left[\sum_{i=1}^{N}\log\Pr(g_i|x_i)-\lambda\sum_{k=1}^{K}\sum_{j=1}^{p}(\alpha|\beta_{kj}|+(1-\alpha)\beta_{kj}^2)\right] \tag{2.5}$$

There are numerous advantages to using shrinkage models, and the elastic net estimation in particular, such as built-in feature selection and more robust and reliable coefficient estimation. This is discussed in more detail by, for instance James et al. [2013, pp. 203–204] and Friedman et al. [2001, pp. 662–663].

Evaluation Techniques The root-mean-square error is one of the measures that can be used for the purpose of evaluating linear regression models: it is defined as the square root of the variance of the residuals between outcome and predicted value and providing the standard deviation around the predicted value, as shown in Eq. (2.6). The advantage over the more general 'mean-square error' (MSE) is that RMSE computes deviations in predictions on the same scale as the data. However, due to the squaring, thus assigning more weight to larger errors, the RMSE is more sensitive to outliers.

$$\text{RMSE}=\sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t-y_t)^2}{n}} \tag{2.6}$$

Additional methods for testing simple associations between variables include correlation analysis. For this either the Pearson correlation coefficient r or Spearmans'ρ can be used. r is computed by dividing the covariance of predictor x and response y by the square root of the product of their individual variances (Eq.(2.7)) or in the case of non-normal distributions the non-parametric Spearmans'ρ as shown in Eq.(2.8) is used. This tests the strength of association between variables by considering their rank. Here, d_i refers to the difference between the ranks of corresponding values x_i and y_i with n being the number of observations.

$$r=\frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} \tag{2.7}$$

$$\rho=\frac{6\sum_i^2 d_i}{n(n^2-1)} \tag{2.8}$$

Having introduced some widely applied techniques in the field, Section 2.7 continues with research in the area of diachronic stylometry.

2.7 Diachronic Linguistic Style

The area of diachronic linguistic analysis is still relatively recent with there only being a few studies to date and methods in need of *refining* and above all *defining* [Stamou, 2007]. However,

long before this, Burrows [1987] conducted an Eigen analysis for a text corpus, including most notably Jane Austen, to identify similarities and differences in authorial style based on the most common, both grammatical and lexical words, finding that the method is able to successfully distinguish between Austen's different writing stages.[11] A similar analysis was considered by Burrows [1989]. This study analysed different literary works using Eigen analysis and again of particular interest was an analysis done on Jane Austen's dialogue based on the most common word types' relative frequencies. Interestingly, when projecting the books' characters (based on their dialogue) onto the new representation, according to Burrows [1989] the chronological dimension is more powerful than the gender dimension, in that 'it is clear that a novelist who was capable of distinguishing the idiolects of her leading characters by gender ceased to enforce the distinction in her later work' [Burrows, 1989, p.311].

In 1987, Burrows initially notes, that there presently is a lack of methods for stylistic comparison, be it between languages of different genre or historical periods and that "the contest of faith between those who believe in a self-generated authorial individuality and those who regard an author as a *tabula rasa* upon which larger cultural forces inscribe themselves" would need to be postponed until these methods are in place [Burrows, 1987, p.61].

2.7.1 Chronological Prediction

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

[11] Eigen analysis is comparable to PCA analysis. 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.

Tabata [1994] considers chronological variation in works of Charles Dickens considering separations by narrative style. Tatata examines third-person narratives from Dickens' early period (all written in the 1830s) and first-person and third-person narrative style written after 1849. In this, narrative style specific words are excluded to focus on subtler differences of stylistic differences. PCA is used to group variance patterns among the 74 most frequent words, which are then projected onto the text samples. This reveals a clear separation between the narrative styles along one dimension and their chronology along the other. For instance, relative pronouns, such as *which* and *who* discriminate more strongly in favour of texts written in 1830s, whereas *that* predominate after 1849. According to Tabata [1994], late Dickensian style also features the adverbial particles *out* and *down*, the pronoun *it*, and the preposition *like*, where some of these shifts have also been noted by other studies with respect to general English first-person narrative. In order to identify the most discriminative chronometers, Tabata [1994] uses a t-test on each marker separately for the third-person narrative samples. The markers *which*, *it*, *out* and *like* are very highly significant ($p < 0.001$). Clustering using only the 21 most discriminative words results in an even sharper distinction between sets.

Another slightly more recent work in the area focuses on detecting changes in writing styles of two Turkish authors, Cetin Altan and Yasar Kemal, in old and new works [Can and Patton, 2004]. Similar to the previous study, works were also separated into two categories of 'young' and 'old' author. Altan's works were sampled from the years 1960–1969 (young) and 2000 (old) and for Kemal one novel each from 1971 (young) and 1998 (old) was selected. For each author, the data was divided into sixteen fixed-size blocks of 2,500 words for each old and new period. Based on this, they considered three different style markers: both type length and token length and the frequency of the most frequent word unigrams. Average type and token length was reported to significantly increase for both authors comparing their old and new works respectively using a t-test. Further, employing different methods, such as linear regression, PCA and ANOVA, they found that word types are slightly better discriminators than type and token length.[12] 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.[13] Analysing usage rates of different type and token lengths using logistic regression showed that word length of three to eight is predominant in Altan's old works, whereas a word length of nine or greater was more representative of his new works. For Kemal, a similar distribution emerges, leading to the conclusion that these results may have been due to Altan's and Kemal's higher mastery of the language later in their lives. Can and Patton [2004] also compare old and new works with respect to characteristic word features, yielding five significant markers for Altan and two for Kemal. Finally, using discriminant analysis, they aim

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

[13] 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.

to find the best chronometer, achieving 98.96% average classification rate for Altan and 84.38% for Kemal, the difference in which they attributed to the greater time distance between Altan's work and consequently the more pronounced development in style. Another literary-motivated analysis considered temporal change in the late 19th century American author Henry James [Hoover, 2007], who is deemed to have changed his style over his creative lifespan [Beach, 1918]. Considering the most frequent word unigrams and a variety of different methods, such as Cluster Analysis, Burrows' Delta, Principal Component Analysis and Distinctiveness Ratio, Hoover investigates natural partitions of James' style into three different temporal divisions of early (1877–1881), intermediate (1886–1890) and late style (1897–1917).[14] These three divisions have also been identified by literary scholars [Beach, 1918]. Furthermore, Hoover notes the existence of gradual transitions, with the first novels of the late period being somewhat different from the rest of them. Analysis of the 100 words with the largest Distinctiveness Ratio that are either increasing or decreasing over time show that James appears to have increased in his use of *-ly* adverbs and also in his use of more abstract diction, preferring more abstract terms over concrete ones.

A property that combines all the previous three studies is analysis based on categorical divisions rather than a continuous analysis of development of style. Especially in the analysis of James' style, the results suggest that he changed his style over time in a fashion that would allow for a continuous analysis of temporal style rather than merely creating broader divisions into different periods as was done in this study. Detecting change between these sub-periods suggests that there might be continuous development of frequency distributions whereby James gradually adopted certain features while at the same time slowly started to abandon others.

2.7.2 Constancy of Style

The work presented by Smith and Kelly [2002] considers a more granular division of works in investigating the question of whether vocabulary richness remains constant over time through examining measures of lexical richness across the diachronic corpora of three playwrights (Euripides, Aristophanes, and Terence). The plays were divided into standardised non-overlapping blocks, each being analysed for certain properties pertaining to lexical richness, such as vocabulary richness, proportion of *hapax legomena* and repetition of frequently appearing vocabulary. Apart from testing constancy of these properties over time, weighted linear regression is used to test associations between these measures and the time of the play's first performance. For this, the property's value in a particular text block is used as the response and the time of performance is used for prediction.[15] The results show that Aristophanes appears to have decreased in his use of *hapax legomena* over time. Interestingly, one of his earlier works, *Clouds*, was

[14] 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.

[15] In order to perform inverse prediction, i.e. predicting the date of an unknown work by the measure, the authors draw a horizontal line at y, with y corresponding to the measures' average in the text and look at the intersection with the estimated regression line.

subjected to redrafting after the first staging, but for which the finishing date is unknown, is predicted to originate towards the end of the playwright's life, indicating that revisions might have been made at a much later stage.

2.8 Conclusion

This chapter presented a review of works in the areas of both stylometry and chronological linguistic analysis. Although there do exist methods in general language change, ageing and pathological language decline as well as stylochronometry, the latter being primarily focused on timeline prediction, methods combining these analyses are somewhat lacking. This may render individual results problematic in that it is not certain which effects can be attributed solely to the individual rather than the other competing influences of background language change or linguistic ageing.

Chapter 3

Data

In this chapter, the data to be analysed is described and motivated. Specifically, this thesis considers both a corpus of American literary authors[1] from the 18th–19th century as well as a reference corpus[2] for American English for the same time period. As part of the analysis in the following chapters, I examine two authors in the corpus in more detail, Henry James and Mark Twain, and section 3.1 motivates this special treatment. The literary authors, presented in Section 3.2, are studied both separately and in unison to illuminate different aspects of style change. Additionally, data preparation is discussed for both the literary authors' corpus and the reference corpus. Section 3.3 considers aspects of part-of-speech processing in more detail, as it forms the basis for the majorities of features extracted as part of this work.

3.1 James and Twain

Works by Henry James and Mark Twain form part of the literary corpus, assembled to investigate individual changes in authorial style over time. However, there is reason to examine them separately from the rest of the literary authors in the corpus. As already discussed in section 2.7.1, James has been subject of analyses of style development and has been deemed to have changed his style in ways that allow for identification of three separate periods: early (1877–1881), intermediate (1886–1890) and late style (1897–1917) [Hoover, 2007]. I have not yet found explicit sources suggesting Twain has undergone similarly measurable change, but a series of dramatic events during his lifetime render him a likely candidate to exhibit change in style. Additionally, several examined sources [Beach, 1918; Canby, 1951] seemed to suggest that it might be interesting to contrast James with Twain given that both authors were highly prominent and prolific authors composing works largely in parallel, yet appeared to have harboured a strong dislike towards each other, that may or may not have resulted in a marked difference in style choices over time.

[1] Hereafter also referred to as: 'literary authors' corpus' or 'LAC'.
[2] Hereafter also simply referred to as: 'reference corpus' or 'RC'.

To begin this investigation of Mark Twain's and Henry James' style over the course of their respective writing careers, this section considers James' and Twain's personal lives to gain insight into their personality, convictions and temperament, since these are likely to also take some manifestation in their writing. Although both authors belonged to the same stream of pioneer American history and are both regarded as having been highly articulate, close observers as well as being very creative writers, at the same time they were also "violently in contrast in temperament, in their art, in their strength and in their weaknesses" [Canby, 1951, p. xii]. Together they are often regarded as representatives of the 'Turn West – Turn East' American, which might explain why "[n]either would or could read the other" [Canby, 1951, p. xii].[3] Twain was often regarded to be offensively American, while James sometimes even shocked his family with his displayed patina of *Britishness*.

In terms of their modes of inspiration and composition, they also widely differed. While Henry James forged his characters by minutiae observation of others, lacking that curious interpretation of self, Mark Twain poured himself into his writings 'transfer[ring] his own life to his creature, who becomes not a study, but a human being realised not merely described' [Canby, 1951, p. 249]. This difference in intimacy with their characters might also be linked to the fact that Samuel Clemens decided to write under the pseudonym *Mark Twain*, thus somewhat enforcing a distance between himself and his characters, whereas this might not have occurred to James, who never got as close to his *subjects*.

3.1.1 Of Mark Twain

~Mark Twain's appeal is an appeal to "rudimentary minds"

HENRY JAMES (ON THE POPULARITY OF MARK TWAIN'S WORKS)

Apart from his career as a writer, Mark Twain appears to have led quite a colourful professional life, trying a variety of occupations, among them printer, river pilot, soldier, secretary, speculator, and miner. In terms of his writing career, he has had a close relationship to his characters, as for instance *Tom Sawyer* represents one aspect of Sam Clemens' self-conflicting personality [Canby, 1951]. Writing under a pseudonym gave Clemens the liberty to create an alter ego, a public character for himself, while distancing himself from it in private as he pleased. This resulted in the creation of two more or less separate identities: Clemens - the sensitive, perceptive friend opposed to Mark Twain - the robust and astringent humorist [Wecter, 1945].

Since his person as well as his life were deeply interwoven with his creative career as a writer, it is worthwhile considering distinctive events in his life. Clemens suffered quite a number of tragedies, especially in his private life, that might partly explain why he seemed to have

[3]Twain could be perceived to be super-normal for life of westward-turning American, while James would abnormal, even for eastward-turning men and women of America [Canby, 1951].

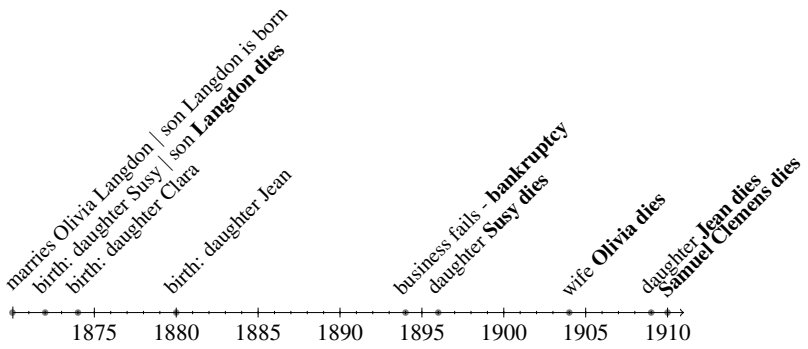


Figure 3.1 – Extract from Twain's life events.

lost faith in humanity in the later part of his life. He lost three of his four children as well as his wife Olivia, with his third daughter dying only one year before him (see figure 3.1). The death of his beloved daughter Susy was a disaster from which he emotionally never recovered [Canby, 1951]. In addition, his business venture in publishing failed and he was practically bankrupt by 1894, which forced him to go on several lecture tours in the years following. Mark Twain's social criticism expressed in Huckleberry Finn (1884) still corresponded "more simply to that of the rational democrat and humanitarian who has not lost faith in the practical effectiveness of the good heart on this earth" [Branch, 1957]. Whereas, later on, he started to shift from confident optimism to bitter cynicism for which the key can be found in his own words in his last travel book *Following the Equator* [Wecter, 1945]:

"EVERYTHING HUMAN IS PATHETIC.
THE SECRET SOURCE OF HUMOR IS NOT JOY BUT SORROW.
THERE IS NO HUMOR IN HEAVEN."

Published in 1889, *A Connecticut Yankee in King Arthur's Court* was the last display of his creative powers in full strength, while *Joan of Arc* (1896) seemed to be a sign of his losing the imagination to carve his own popular characters and rather choosing historical figures as basis for writing [Canby, 1951].[4] The *Mysterious Stranger* (1908) was left unfinished – it might have been perceived as 'too strong' and thus not combinable with the goal of popularity [Canby, 1951].

3.1.2 Of Henry James

"Once you put it down, you simply can't pick it up!"

MARK TWAIN (ON A BOOK BY HENRY JAMES)

[4] See appendix A: Table A.15 for the detailed list of Twain's publications included in the literary authors' corpus.

In contrast to Twain, James seems to have lived a somewhat secluded life, where while growing up, he studied primarily with tutors and was relieved from serving in the civil war owing to a bad back. In some sense, he remained a spectator of life (never had his own family), while observing and writing abundantly over his lifetime. James is widely regarded to have evolved dramatically in terms of his style and later distanced himself from earlier works to an extent of not admitting them to a collective edition. In addition, he later revised numerous early works that after his changes were considered closer to his later style.[5]

In terms of composition, James built novels primarily upon a motive or idea, similar to George Meredith or George Eliot [Beach, 1918, p.11]. Further, his objective with respect to his readers was to "make [them] acquainted with the characters of the *dramatis personae*" rather than "make us see his figures" like Charles Dickens [Beach, 1918, p.13]. There is less characterisation through dialogue and his characters are very little differentiated through what they express, and come alive more through his description. Similarly, James was strongly imbued with the principle of 'art for art's sake' [Beach, 1918, p.30] and even though generally moral, he showed himself detached and neutral and did not pass judgement onto his characters. Thus, "[t]he necessity of an unconscious morality, accepted, not argued about, was the basis of Henry James' writings all his life" [Canby, 1951, p.46].

James' late style is dominated by an unusual and intricate syntax and what has been referred to as "free, involved, unanswered talk" [Schilleman, 2013]. William James, observed in a letter to his brother in 1902 that in his view Henry had "reversed every traditional canon of story-telling"(especially the fundamental one of telling the story)"[Smith, 2012, p.6]. It is worth noting that Mark Twain and William James (Henry's brother) maintained an active friendship throughout their lives, both being interested in the psychical research and paranormal phenomena.[6]

However, in spite of apparently very different modes of composition, both Henry James and Mark Twain appeared to have shared an avid interest in history. Both Blair [1963] and Walsh and Zlatic [1981] note that history played an important part in Twain's personal as well as his professional life, even if he did not always incorporate his knowledge in his works consistently [James D. Williams, 1965]. As for James, in his 1884 essay *The Art of Fiction* he claims his place among the historians, since as a novelist he chronicles life and as "picture is reality, so the novel is history" [James, 1884].

James and Twain appear to occupy a particular position in the public mind and given that both were prominent and prolific writers, possibly representing extreme opposites renders them interesting candidates to contrast. However, the fact that they have been influential writers that drew considerable interest from literary scholars and linguists does not necessarily mean that their style change is different from or more pronounced than other contemporary authors that have received less attention. This work aims to investigate in part whether James and Twain

[5]For this analysis, only the earlier versions of these works were collected.

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

evolved differently from other comparable writers in terms of style or whether there is possibly a discrepancy between what is publicly perceived and what can actually be detected. The next section introduces the remaining literary authors that James and Twain are compared to as well as the reference corpus data.

3.2 Data Sets

The data analysed for this research is divided into two main sets: twenty-two literary authors ranging from 1847 as year of first publication to 1923 as year of last publication. The reference corpus provides a background language corpus spanning 1830–1929. Although the data was collected with no conscious prejudice, no data collection can be truly unbiased. Some of the biases that could conceivably have influenced the selection of the particular data samples used are discussed in the following, while others might remain buried in the data. Not all biases are necessarily going to be relevant but should be mentioned nevertheless. The results of the analysis are specific to this data set and can therefore not provide sole evidence for developments in language change, but should rather be seen to contribute to the collective evidence. In particular, Section 3.2.1 considers aspects of the collection of literary authors, followed by a discussion of the reference corpus in Section 3.2.2.

3.2.1 Literary Authors

Table 3.1 shows the set of literary authors, comprising twenty women and twenty-two men, all of whom composed work between 1847–1923.[7] The corpus was populated in the following way: including Mark Twain alongside Henry James has already been motivated in the previous section. The remaining authors were chosen by first assembling a list of male and female American authors of the 19th–20th century using *Wikipedia*[8] and then selecting a subset of this list of authors, all of who had a few works publicly available and spread out over at least twenty years. Also, for the purpose of estimating stable word distributions, it was decided that works had to be at least 150 kilobytes in length thus discarding authors with multiple shorter works. Consequently, there might be a bias towards more prominent writers, as there could have been more incentive to make their data publicly available. For instance, this may result in a shift towards only certain words or expressions being used more frequently throughout. Also, there is little to no racial diversity in the data set as all authors were white, and even though individuals, such as Harriet Beecher Stowe described African Americans' conflicts in her writings, most authors probably remained in their sphere and wrote predominantly about the type of society they themselves were exposed to. Therefore, any inferences based on this set

[7] The corpus is described and motivated by Klaussner and Vogel [2018a]. The data set is available at `www.scss.tcd.ie/clg/DCLSA/` – last verified March 2018.

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

Table 3.1 – 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
Alice Brown	1884–1922	F	12	5.7	1064566
Amanda Minnie Douglas	1866–1914	F	51	24.5	4500421
Constance Fenimore Woolson	1873–1895	F	12	6.7	1204937
Edith Wharton	1897–1920	F	10	3.5	609351
Elizabeth Stuart Phelps Ward	1866–1907	F	21	5.8	1055611
Gertrude Atherton	1888–1923	F	19	9.1	1628163
Harriet Beecher Stowe	1852–1886	F	18	11.2	2049014
Louisa May Alcott	1854–1893	F	16	5.6	1027950
Marion Harland	1854–1914	F	15	9.0	1572983
Susan Warner	1850–1884	F	29	18.6	3467028
Charles Dudley Warner	1872–1899	M	14	6.1	1088452
Edgar Saltus	1884–1919	M	17	3.6	650825
Francis Marion Crawford	1882–1908	M	41	23.3	4238660
Harold McGrath	1903–1922	M	15	5.3	945365
Henry James	1877–1917	M	32	17.3	3123582
Horatio Alger jr	1866–1906	M	37	10.3	1840445
Mark Twain	1869–1916	M	23	11	1990085
Robert W. Chambers	1894–1922	M	38	20	3465933
Timothy Shay Arthur	1847–1890	M	30	10.7	1933432
Upton Sinclair	1898–1922	M	17	8.6	1572977
William Dean Howells	1867–1916	M	38	16.7	3063271
William Taylor Adams	1855–1896	M	49	17.5	3208971

of literary authors does not necessarily extend to the population of American literary authors at large. First temporal alignment in the corpus and inter-author relationships are described, followed by a discussion on aspects of data collection, error correction and analysis.

Author Relationships 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).[9] Elizabeth Stuart Phelps Ward and Amanda Minnie Douglas both began writing about 15 years later in 1866. The remainder of the female authors' first contribution is somewhat spread out: Constance Fenimore Woolson (1873), Alice Brown (1884), Gertrude Atherton (1888) and lastly Edith Wharton (1897). As for the male authors, Charles Dudley Warner, Mark Twain, William Dean Howells and Horatio Alger jr also made their first appearance within a few years of each other (1866–1872). The second big wave of male authors' first publication clusters around the 1880s: Henry James (1877), Francis Marion Crawford (1882), and Edgar Saltus (1884). Timothy Shay Arthur and William Taylor Adams started publishing slightly earlier than the rest, 1847 and 1855, respectively, and both remained active for about 40 years. Thus, these earlier timelines still have considerable overlap with most of the other writers in the corpus. An exception to this are Upton Sinclair, Robert W. Chambers, as well as Harold McGrath and Edith Wharton, who only started their career in the 1890s or beginning of the 20$^{\text{th}}$ century. However, most authors in this corpus should be comparable in that they composed work over at least 20 years in parallel.

Apart from Henry James' and Mark Twain's rivalry described in more detail in section 3.1, there existed some other connections between authors of this corpus. Mark Twain and Charles Dudley Warner wrote *The Gilded Age* together.[10] Elizabeth Stuart Phelps Ward seems to have been an admirer of Harriet Beecher Stowe and referred to her in 1896's *Chapters from a Life* as the "greatest of American women". Constance Fenimore Woolson, a grandniece of James Fenimore Cooper, quoted William Dean Howells in one of her works and established a friendship with Henry James. Her 1884 *East Angels* is seen as a response to James' *Portrait of a Lady* [Kreiger, 2005]. Susan Warner's 1850's *Wide, Wide World* has been described as a "Feminist *Huckleberry Finn*".[11]

[9] 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.

[10] As Twain is listed as first author, it is assigned to his corpus. There is a trade-off between adding more data points to make the stylistic analysis more stable and not having stylistic interference from collaborative works. Ideally some individual analysis situating the work as closer to either one of the two writers should precede a temporal analysis or in cases where it does not cluster well with either, it should excluded entirely. Collaborations are the exception here and were added because of data sparsity judging that it may be more beneficial to include them even if their inclusion might slightly distort the results.

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

Table 3.2 – Common OCR errors and their correct possible realisations, their raw counts and % of processed IA tokens.

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

Data Collection The set of literary authors was mainly collected from *Project Gutenberg (PG)*[12] and supplemented with works from the *Internet Archive (IA)*.[13] Project Gutenberg is the more desirable source given that the data is hand-transcribed rather than scanned automatically. However, in this case acquiring data with a time stamp close to the first publication date was essential and for this reason and especially when the equivalent Gutenberg version did not have a time stamp, the Internet Archive version was chosen instead if available. The Internet Archive contains scanned versions of books using Optical Character Recognition (OCR), and the quality of the processing varied considerably across books and sponsors. In this a trade-off had to be found, balancing accurate time stamp and quality of processing. Occasionally, when content was very noisy due to OCR errors, files were not included at all. In all cases, the date of a file was decided by taking the first available date, e.g. first copyright or publication date, unless a preface clearly stated that the work had been subject to explicit revisions. The issue with dating in this case is that both dating a work too early or too late would distort the results. All data was prepared for processing by manually removing parts that were written at a different time from the main work or introductions or comments not by the author, such as notes or introductions by editors. Additionally, table of contents were also removed, as these do not usually follow a normal sentence structure. Minimal preprocessing was needed for PG files, but the books sourced from the IA could be rather noisy, and as upon inspection each file appeared to have different types of OCR errors, it was deemed best to process each file manually to correct scanning errors and remove unwanted formatting sequences. One of the issues with automatically correcting these errors was that even within one file, a misread character could refer to multiple different correct character realisations and only manual examination of the context could accurately determine the correct realisation.[14] Errors that had only one pos-

[12] `http://www.gutenberg.org/` – last verified February 2018.
[13] `https://archive.org/` – last verified February 2018.
[14] ‘Character’ here refers to alphanumeric letter.

sible correct version could be corrected using regular expressions, but manual correction was necessarily in cases where there was more than one possible correct version, e.g. the error *011* could correspond to both *no* and *on* even within the same file.

Table 3.2 shows some of the most common OCR errors and their possible correct realisations as well as occurrence of these and their rates as percentages of the raw corrected tokens in IA texts. All whitespace-separated items in the raw texts add up to 14140296 tokens, which reduce to 13614013 tokens in the manually processed version (a reduction of 4%). I estimated the number of broad differences between the two versions by considering the lines changed compared to all lines in the processed version, i.e. 137594/2146720=0.064 (6.4%).[15] It is important to note that there could be multiple changes per line and simple deletion of superfluous headings or page numbers would not be as time-intensive as manual correction of OCR errors. All processed works add up to 554 files in total, 400 (176.9 MB) from Project Gutenberg and 154 (73.7 MB) from the Internet Archive. When reducing the set to unique *author-publication year* combinations, 409 cases were left. Generally, all counts for a particular feature in a year are joined and relativised as one, even if originating from different books as long as these were written in the same year. For this, features are first extracted from each text and then added rather than concatenating text for the reason that for longer word sequences, one would then also count ngrams between two separate novels that were never actually written.

3.2.2 Reference Corpus

The reference corpus is meant to provide samples of *representative* language usage for each year that are unlikely to be influenced by age or individual stylistic patterns as the authors within each year presumably change from one year to the next. The main motivation for analysing individual stylistic changes with respect to background language change is the disentanglement of individual changes due to age or style change from those that belong to underlying language change and which should not be attributed to the individual author. For instance, one might discover a decrease in usage of a particular feature over time for Mark Twain; yet without knowing if the same feature did not decrease in usage in general, one cannot know whether this is noteworthy at all. While both the individual and the general language change analyses are of interest in their own right, they also assign meaning to the respective other, indicating whether particular events are likely to be unusual.

The reference corpus was assembled by taking an extract from the *The Corpus of Historical American English (COHA)* [Davies, 2012].[16] COHA is a 400-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

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

[16] A limited free version is accessible on: http://corpus.byu.edu/coha –last verified March 2017.

from fiction, popular magazines, newspapers and non-fiction books, which are again balanced across sub-genre, such as drama and poetry.[17] The COHA data was compiled from different sources, some of which were already available as part of existing text archives (e.g., *Project Gutenberg* and *Making of America*), whereas others had to be converted from PDF images to text or scanned from printed sources. The corpus allows for different levels of linguistic change analysis, i.e. lexical, morphological, syntactic and semantic.

While the reference corpus is less externally biased as it contains copyright pieces alongside freely available data, it might be internally biased, e.g. through editorial decisions if, for example, several newspapers were owned by the same person. For instance, by the mid-1920s, the businessman William Randolph Hearst had acquired 28 newspapers, that might have been subject to the same editorial decisions, distorting one's perception of what language was representative for the time. The newspaper collection of this corpus does not appear to contain a Hearst-owned newspaper, however, three of his magazines, the *Cosmopolitan*, *Good Housekeeping*, and *Harper's Bazaar* are among the list of magazines for this corpus.[18] In any case, even without direct influence through a high proportion of newspaper samples in the corpus, the fact that a quarter of the American people read a Hearst newspaper at the time could potentially have given one man the means to somewhat influence language usage and change.[19]

[17] There is an excel file with a detailed list of sources available on: `http://corpus.byu.edu/coha/` – last verified March 2018.

[18] `http://en.wikipedia.org/wiki/Hearst_Communications` – last verified: March 2018

[19] `http://www.encyclopedia.com/people/literature-and-arts/journalism-and-publishing-biographies/william-randolph-hearst` – last verified: March 2018

Table 3.3 – Part-of-speech tags used in the Penn Treebank Project

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

3.3 Part-of-Speech Tagging

The construction of part-of-speech (POS) tags forms the basis of various features used in the following analyses. Apart from simple POS features, syntactic word features as well as tense features also rely on POS tagging for feature extraction.

The majority of the tagged data is prepared using the *TreeTagger* through the *koRpus* package in R [Schmid, 1994; Michalke, 2014].[20] The exception to this is the extraction of tenses in Chapter 4, where a different tagger was used instead, i.e. the *OpenNLP POS Tagger*, employing the *NLP*[Hornik, 2016] and *openNLP*[Hornik, 2015] packages. The switch was made to facilitate alignment between POS and chunk tags. Chunk tags were used to extract verb tenses, thus avoiding an expensive syntactic parse tree analysis that would otherwise have been necessary. Both taggers follow the *Penn Treebank* tagset (shown in Table 3.3) and report high accuracies (96–97%) [Schmid, 1994]. The TreeTagger differs from other conventional ngram taggers in that it estimates transition probabilities with a decision tree. The OpenNLP POS Tagger uses a Maxent (Maximum Entropy) model to predict the correct POS tag out of the tag set, also taking into account low frequency features. The second parser's output is only employed indirectly to classify tenses rather than directly by using the POS tags as in the case of POS features or syntactic word features. In all cases, the data was tagged automatically, verifying correctness of tags on random samples, although this was not analysed quantitatively.

Tagging depends somewhat on context, for instance most determiner word forms are unambiguous, such as *the* or *a*, as there cannot be multiple occurrences in the same noun phrase. However, *all*/*both* can be both a simple determiner as well as a predeterminer. The TreeTagger appears to be well able to distinguish those latter cases, while it deviates from the manual in those cases, where *all*/*both* are used pronominally without a head noun. There are a few isolated cases, where these are not tagged as determiners according to the manual but as proper nouns instead. According to the Penn Treebank Tagset manual[21], the different categories are defined as follows:

- **Adjectives** (JJ, JJR, JJS):

  Hyphenated compounds, such as *one-of-a-kind* as well as ordinal numbers, such as 'fourth-largest' that are used as modifiers are tagged as JJ. Comparative and superlative simple forms ending in *-er* and *-est* respectively are tagged as JJR and JJS. *More* and *less* and *most* and *least* are tagged as JJR and JJS respectively when occurring on their own. If adjectives do not have the comparative or superlative ending, even when having a comparative or superlative meaning, these are simply tagged as JJ.

[20] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

[21] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Penn-Treebank-Tagset.pdf`

- **Adverbs** (RB, RBR, RBS, WRB):

  This category includes most words ending in *-ly*, as well as words signifying degree, such as *quite*, *too*, *very*, post-head modifiers as in *good enough* and markers of negation like *not*, *n't* and *never*. Only adverbs ending in *-er* with a strictly comparative meaning should be tagged as RBR (RBS for superlative). Wh-adverbs, such as *how*, *where*, *why* and *when* in their temporal sense are tagged as WRB.

- **Cardinal number** (CD):

  This label is used for references to numbers in different formats, e.g. *3* and *three*.

- **Conjunction** (coordinating (CC), subordinating (IN)):

  Coordinating conjunctions cover expressions, such as *and*, *but*, *nor* as well as mathematical operators *plus*, *minus*, *less*. Also included is *for* when used in the sense of *because*. *So* used in the sense of *so that* is tagged as a subordinating conjunction (IN). Subordinating conjunctions are conflated with prepositions, however still distinguishing between a preposition that precedes a noun or prepositional phrase, and a subordinating conjunction that precedes a clause. The preposition *to* receives its own tag: TO.

- **Determiner** (DT, PDT, WDT):

  The DT label includes articles *a(n)*, *every*, *no*, *the* as well as the infinite determiners *another*, *any*, *some*, *each*, *either*, *neither*,*that*, *these*, *this*, *those* and instances of *all* and *both* that do not precede another determiner or possessive pronoun. All other instances are tagged as predeterminer (PDT). Any noun phrase can contain at most one determiner, rendering tags for the same lexical representation context dependent, for instance *such* as in *the only such case* is an adjective whereas it should be tagged as a predeterminer in *such a good time*. The WDT tag covers *which* as well as *that* when it is used as a relative pronoun.

- **Exclamation** (UH):

  this covers expressions or fillers, such as *oh*, *please*, *see* or *yes*.

- **List item marker** (LS):

  This tag includes letters or numerals used in lists.

- **Modal verb** (MD):

  This type includes all verbs that do not take the *-s* ending in the third person singular present: *can*, *could*, *dare*, *may*, *might*, *must*, *ought*, *shall*, *should*, *will*, *would*.

- **Nouns** (NN, NNS, NP NPS):

  Nouns are divided into common (NN, NNS) and proper (NP, NPS) nouns with *S* signifying plural in each case.

- **Particle** (RP):

  This category covers mostly monosyllabic words that also double as directional adverbs and prepositions.

- **Personal pronoun** (PP, PP$, WP, WP$):

  The PP tag covers personal pronouns proper irrespective of case as well as reflexive pronouns ending in *-self/-selves* and the nominal possessive pronouns *mine*, *yours*, *his*, *hers*, *ours* and *theirs*. Adjectival possessive forms (*my*, *your*...) are tagged as PP$. Wh-pronouns, such as *what*, *who*, *whom* are tagged as WP. Possessive wh-words, such as *whose* are tagged as WP$.

- **Possessive ending** (POS):

  For this, noun's possessive endings *'s* or *'* are split off and tagged as if they were a separate word.

- **Symbol** (SYM):

  This type should be used for mathematical, scientific and technical symbols that are not words of English.

- **Existential** (EX):

  This tag applies to occurrences of the unstressed *there* that triggers inversion of the inflected verb and the logical subject of a sentence.

- **Verbs** (VB, VBD, VBG, VBN):

  The tag VB includes imperatives, infinitives and subjunctives. VBD covers verbs in the simple past tense. VBP and VBZ both apply to present tense, with VBZ referring to 3rd person singular and VBP to everything other than 3rd person singular. VBG is used for gerund or present participle and VBN to mark past participle.

The different tags and their functions are summarised in Table 3.3.

3.4 Conclusion

This chapter introduced the data sets used for the linguistic style analysis, i.e. the literary authors' corpus and the corresponding reference corpus. In particular, Section 3.1 provided some motivation to consider James and Twain more specifically as part of the following analysis. Section 3.2 discussed aspects of the two data sets including their preparation for feature extraction. Finally, Section 3.3 presented aspects of part-of-speech tagging that form the basis for the majority of features used in this research.

Chapter 4

Linguistic Ageing in Literary Careers

This chapter considers linguistic change within the ageing dimension and investigates whether and to what extent general language change may interfere with effects previously solely attributed to ageing. After having examined general language change patterns with respect to six linguistic variables previously associated with ageing, a general model for measuring the effects of ageing in the literary authors (introduced in Chapter 3) is proposed taking into account the background language at the same time. This chapter therefore not only investigates linguistic ageing with respect to literary authors, but also introduces a method to disentangle background language change from possible ageing effects.

Specifically, Section 4.1 briefly discusses the original study in linguistic ageing by Pennebaker and Stone [2003] as well as motivations for the current work. Section 4.2 presents feature extraction methods and statistical models. Section 4.3 then examines change in the background language and proposes a model taking into account this influence when analysing individuals with respect to ageing; Section 4.4 then discusses the results. This chapter closes with Section 4.5, a summary of the results.

4.1 Introduction

This research is an investigation into aspects of language change over some literary authors' life spans. An individual's language change is subject to not only his or her own stylistic development as a writer, but also to the general underlying language change all individuals of a particular linguistic community are exposed to, as well as linguistic changes that are possibly due to ageing, as previously discussed in Section 2.5.2. In order to examine stylistic changes over time accurately, one has to control for two different types of possible confounding factors: the influence of general underlying language change pervading the language of the entire reference population and effects of ageing as part of language development. Not accounting for these factors may lead to misinterpreting effects in the individual as stylistic when they may be due to more systematic alternative factors. Specifically, effects due to ageing may be

erroneously interpreted as mere stylistic changes in the individual author's canon, unless other comparable authors are examined in parallel.

In order to clearly identify effects due to ageing, a number of conditions have to be met, i.e. effects have to be present in multiple authors irrespective of exact time period, while in each case controlling for the background language at the same time. If no age-aligned effects across authors can be found, there is little evidence for the influence of ageing on language in literary authors. When age-aligned effects among authors are present while having controlled for background language change, this still has to be investigated with respect to common stylistic changes, especially when authors wrote largely in parallel. To somewhat exclude this possibility, authors from different time periods should be compared, so that they align differently based on their age than the publication date of their works. Especially, if prediction based on *age at time of publication* is outperformed by simple *time of publication*, there is little evidence for common ageing effects. Another important precaution is to conduct further replication studies of this kind to examine these effects across different types of language, e.g. spoken as well as written text.

Table 4.1 – 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**	**Author project**	**Disclosure project**	
			Experimentals	*Controls*
Social and identity				
First-person singular	*I, me, my*	$-0{\cdot}26^{*}$	$-0{\cdot}18^{*}$	$-0{\cdot}18^{*}$
First-person plural	*we, us, our*	$0{\cdot}03$	$-0{\cdot}01$	$-0{\cdot}27^{*}$
Time orientation				
Past-tense verbs	*was, went, ate*	$0{\cdot}08$	$-0{\cdot}20^{*}$	$-0{\cdot}22^{*}$
Present-tense verbs	*am, see, goes*	$0{\cdot}09$	$0{\cdot}05$	$0{\cdot}03$
Future-tense verbs	*will, shall*	$0{\cdot}22^{*}$	$0{\cdot}19^{*}$	$0{\cdot}10^{*}$
Cognitive complexity				
Big words (> 6 letters)	*pontification*	$0{\cdot}10$	$0{\cdot}35^{*}$	$0{\cdot}36^{*}$

Table 4.2 – P&S's 'Characteristics of Authors Chosen for the Author Project' [Pennebaker and Stone, 2003, p.297]

Author	**Nationality**	**Sex**	**Life span**	**Productive years**	**Genre**	**Analyzed works(n)**	**Words per work (M)**	**Aging coefficient correlation**
Louisa May Alcott	United States	F	1832–1888	1854–1886	Novels, stories	19	40,273	−0.05
Jane Austen	England	F	1775–1817	1787–1817	Novels, stories	13	68,120	0.23
Joanna Baille	Scotland	F	1762–1851	1789–1827	Plays	20	18,921	0.60**
Charles Dickens	England	M	1812–1870	1836–1870	Novels	15	257,777	−0.23
George Eliot	England	F	1819–1880	1859–1876	Novels, stories	10	157,751	0.63*
Robert Graves	England	M	1895–1985	1910–1975	Poetry	100	1,689	0.18†
Edna St. Vincent Millay	United States	F	1892–1950	1917–1947	Poetry	21	3,850	0.72**
William Shakespeare	England	M	1564–1616	1591–1613	Plays	37	22,975	0.03
William Wordsworth	England	M	1770–1850	1785–1847	Poetry	64	6,074	0.37**
William Butler Yeats	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 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. F = female; M = male. †: $p \leq 0.08$ *: $p \leq 0.05$ / **: $p \leq 0.001$

The present study extends research by Pennebaker and Stone [2003] (hereafter also: P&S), who investigated how the age of a person affects certain linguistic categories, such as preference for particular pronouns or tenses with respect to two very different data sets: one based on self-reports from emotional disclosure studies (the 'Disclosure project'; hereafter: 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). In this, they examined both simple linear and quadratic effects for the Disclosure project.[1]

Table 4.1 shows P&S's results for individual age-variable correlations for both data sets (limited to those variables that are analysed as part of the current research, as unfortunately, this work only offers 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, present and future tense, and 'big words' all three correlations are in the same direction, although only in two cases all of them are also significant.

Table 4.2 shows the collection of authors in the AP of the P&S study. It is balanced across genders, but contains some idiosyncrasies, such as that most authors originated from Great Britain (England and Scotland), except for Louisa May Alcott and Edna St. Vincent Millay, who were from the United States. Genre types are spread across novels, plays and poetry, a fact that could present a confounding factor for the analysis of pronouns that are usually distributed somewhat differently in these text types. The relevant issue in this context is that authors' works are spread across three centuries and language use would be expected to somewhat vary between the 16^{th} and 20^{th} century. 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 more 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 4.2 shows the result of using regression weights for the LIWC variables based on the Disclosure data to create an ageing coefficient for each individual author, which was then correlated with age. Thus, larger correlations signify more similarity to the previous studies 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 of the two data analyses appears to have considered or evaluated the influence of general language change on the text samples in question, something that is remedied as part of this work.

[1] This is discussed in more detail as part of Section 4.3.

Previous design As already discussed in Section 2.5.2, P&S [2003] based their analysis on the LIWC system, whose categorisation scheme is not openly accessible. This renders replication of less objective linguistic variables, such as *negative* or *positive* emotion words rather difficult, not taking into account the fact that these may be somewhat context dependent.[2] In the case of first-person singular/plural, it was not specified whether non-reflexive pronouns were included in spite of not being listed among the examples in Table 4.1. The examples for tenses also raised questions, in particular it is not clear whether and how *aspect* in tense was treated in their analysis, as the table only lists examples of the main tenses without explaining how other types, such as *present perfect* or *future perfect* are classified. Originally, P&S [2003] also included, what they refer to as 'time-related' words, such as *clock*, *hour* and *soon*. One can assume that this list also includes temporal adverbs in general like *yesterday* or *today*. These expressions may change the interpretation of regular tenses and excluding them could create shifts between tenses. However, this may not be a trivial problem, as sometimes the overall tense would be more strongly signalled 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.*

This suggests the need for a more intricate study of related cases and design of an appropriate classification system. As this could not be done justice as part of the present work, only verb tenses are used to approximate the overall tenses. The main effect of not including temporal adverbs may result in a shift from present to future tense counts.

The following analysis considers several aspects to examine the question of linguistic ageing in literary authors with respect to background language change.[3] The main issue to investigate is whether the variables listed in Table 4.1 show significant development over time that could lead to misinterpretation of the same variables in the individual authors' case. In this, I compare to both their results for the AP and and the DP. The former would arguably be more similar to the corpora underlying the present research, but occasionally showed more inconclusive results for the variables considered here.

In the following, Section 4.2 outlines the methods that form the basis for the background language experiments and more comprehensive modelling of language effects in Section 4.3.

[2] To the best of my 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.

[3] A related pilot study has been reported on by Klaussner and Vogel [2017].

4.2 Methods

Section 4.2.1 describes how features were extracted, followed by section 4.2.2: the statistical models used for the analysis.

4.2.1 Feature Extraction

and the corresponding background language for the same time period was used, also hereafter referred to as 'reference corpus' or 'RC'. After preparing the text as described in Section 3.2.1, the relevant features needed to be extracted for all texts in the corpus.[4] Cognitive complexity as measured by the number of long-letter sequences over all tokens was the simplest, least ambiguous feature to extract and was computed by counting the number of words whose word length was equal or greater than six. For extracting first-person singular and plural pronouns the word in conjunction with the part-of-speech tag were used to identify the correct items, e.g. to avoid uses of *I* that refer to numbering. As my experiments did not reveal differences between including or excluding reflexive pronouns, this analysis only reports on non-reflexive pronoun types.

The most difficult feature type to define was *tense*, which incidentally also required more intricate design for accurate identification: 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*. Still somehow items, such as *I'm going to school* had to be distinguished from *I'm going to go to school*.

In order to be able to make these distinctions, I designed a program that uses chunk tags to extract verb phrases and then analyses 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 tense* type includes: *simple present*, *present progressive*, and conditional and modal variants, such as *can/ could/ may* go. The *past tense* type captures *simple past*, *present perfect*, *past perfect*, *past progressive* and similar to the *present tense* 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*.

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 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

[4] For 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.

a categorical variable indicating the *author* (A) of a text.

4.2.2 Statistical Modelling

This section describes some aspects connected to the statistical analysis, i.e. regression models and standardisation techniques, followed by model assessment. This builds on methods introduced earlier as part of Section 2.6.

Regression models

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 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 frequency over time. The simplest model in this context only has an intercept β_0 and predictor x as well as random error ε as shown by Eq. (4.1). A quadratic model would add another predictor $\beta_2 x_i^2$ to this as exemplified by Eq. (4.2).

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{4.1}$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \tag{4.2}$$

However, when analysing the LAC, 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 be (also) 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 change over time instead.

For the literary authors, random effects models were used to account for individual variation between authors, as shown by Eq. (4.3), where y_{ij} is the response variable for author j at age i (year i), T_{ij} is a placeholder for the predictor *age* or *year* and represents the individual-specific random effect. A_j is the author-specific random effect and ε accounts for the error term. Similarly, Eq. (4.4) shows the the same for the quadratic model, adding predictor $\beta_2 T_{ij}^2$.

$$y_{ij} = \beta_0 + \beta_1 T_{ij} + A_j + \varepsilon_{ij} \tag{4.3}$$

$$y_{ij} = \beta_0 + \beta_1 T_{ij} + \beta_2 T_{ij}^2 + A_j + \varepsilon_{ij} \tag{4.4}$$

Table 4.3 – 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 Spearmans' ρ for departures from linearity; this is indicated by a superscript ρ. Significance levels are indicated by: *: $p \leq 0.05$ / **: $p \leq 0.01$ / ***: $p \leq 0.001$.

LIWC Variable	**P&S: Disclosure Project**			**P&S: Author Project**	**Reference Corpus**		
	r	β	β^2	r	r	β	β^2
Social and identity							
First-person singular	$-0{\cdot}13^{**}$	$-0{\cdot}14^{**}$	$-0{\cdot}2$	$-0{\cdot}26^{*}$	$0{\cdot}18^{\rho}$	$0{\cdot}0005^{!}$	$0{\cdot}000003^{!}$
First-person plural	$-0{\cdot}12^{**}$	$-0{\cdot}13^{**}$	$0{\cdot}19^{**}$	$0{\cdot}03$	$-0{\cdot}60^{\rho***}$	$-0{\cdot}0006^{!***}$	$0{\cdot}0000008^{!}$
Time orientation							
Past-tense verbs	$0{\cdot}04^{**}$	$-0{\cdot}16^{**}$	$0{\cdot}01$	$0{\cdot}08$	$0{\cdot}7^{\rho***}$	$0{\cdot}004^{!***}$	$-0{\cdot}000002$
Present-tense verbs	$-0{\cdot}02$	$0{\cdot}04^{*}$	$0{\cdot}06^{**}$	$0{\cdot}09$	$0{\cdot}16$	$0{\cdot}0003^{!}$	$0{\cdot}000005^{**}$
Future-tense verbs	$0{\cdot}00$	$0{\cdot}14^{**}$	$-0{\cdot}02$	$0{\cdot}22^{*}$	$0{\cdot}04$	$0{\cdot}00001^{!}$	$0{\cdot}0000002$
Cognitive complexity							
Big words (> 6 letters)	$0{\cdot}13^{**}$	$0{\cdot}26^{**}$	$-0{\cdot}03$	$0{\cdot}10$	$-0{\cdot}51^{\rho***}$	$-0{\cdot}02^{!***}$	$-0{\cdot}000004$

For fitting linear and normally distributed models, the R package: *nmle* 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].

Ideally, in order to uncover true effects of ageing, there has to be a common model for all literary authors using their *age* rather than the *publication year* of their texts as predictor. Otherwise, it is more likely that effects are to be attributed to general language change or other stylistic changes.

Standardisation Before computing regression models, the predictors *age* and *year* were standardised. In order to preserve similarity with P&S's study with respect to ageing models, the data was standardised by either computing z-scores, i.e. subtracting the mean and dividing by one standard deviation for simple linear regression models or taking the absolute value of the difference from the mean over the sample for the quadratic models.

Assessing Model Assumptions For evaluation, either Pearson correlation coefficient r or Spearmans' ρ were used, for normally and non-normally distributed data respectively. The decision 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 kurtosis, skewness, nonlinear link function (for testing linearity) and heteroscedasticity.[5]

4.3 Experiments

This section begins by examining background language change with respect to the six linguistic variables outlined in Table 4.1. Having considered background language change, Section 4.3.2 then investigates how these effects can be explicitly modelled in the case of the literary authors. This also allows one to determine to what extent background language may be responsible for effects observed in the individuals.

4.3.1 Background Language Change

Examining the change in linguistic variables over time raises the question to what extent these variables were subject to general language change, especially when considering a time span of $\sim$40 years or more. In order to be able to clearly attribute change to either stylistic or ageing factors requires ensuring that the variables in question did not undergo significant language change, for instance in the form of a change in trend either decreasing or increasing over time. To be able to assign meaning to measures of linguistic ageing, an analysis of the change in the background language is addressed as part of this section.

[5] Computed through the *gvlma* package in R [Pena and Slate, 2014].

Table 4.3 shows correlation results for the reference corpus, P&S's Disclosure project and P&S's Author project. The results for computing linear regression models for both simple linear (β) and quadratic models (β^2) are displayed only for the reference corpus alongside P&S's DP as the same model computations were not available for P&S's AP. This reference corpus shares characteristics with both of P&S's studies in that it covers a similar length of time span as the P&S's DP ($\sim$ 70 years) and years contain multiple individual samples rather than a strict within-subject design as was used in the case of their Author project. However, the data in the reference corpus is continuous in that it is genuinely sampled from different time periods, which makes it more comparable to P&S's AP (or at least for some of the authors within) than to the P&S's DP, as some of their data representing different age groups could originate from the same time period. Therefore, exact comparison between either pair is difficult and results have to be viewed with caution. Also, as linear models were not available for the Author project, so comparisons between the AP and the current work can only be made on the basis of the correlation analysis results.

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 results. 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 first-person plural[6] pronouns and past tense, where the effect is in the same direction as for P&S's DP and long-letter sequences, where effects are in the opposite direction for both of P&S's studies..

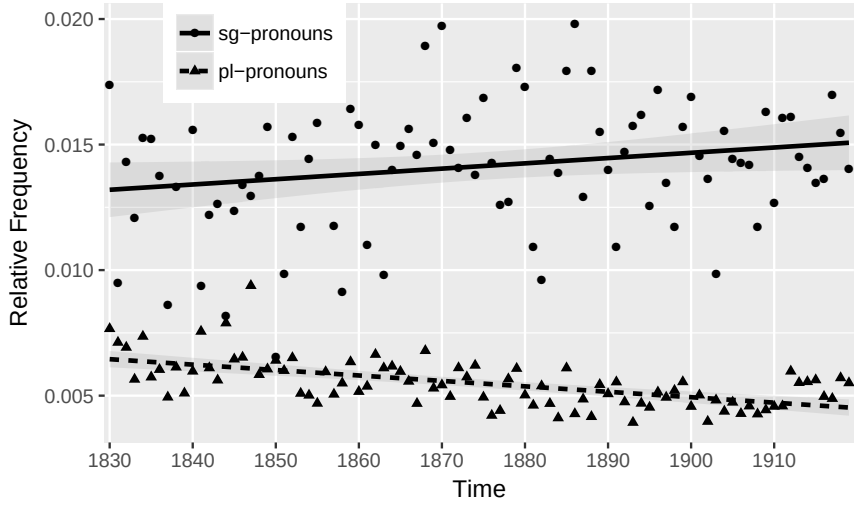
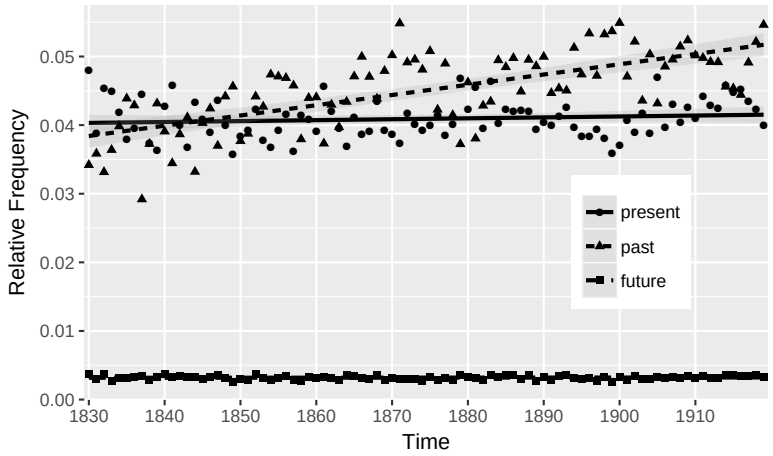


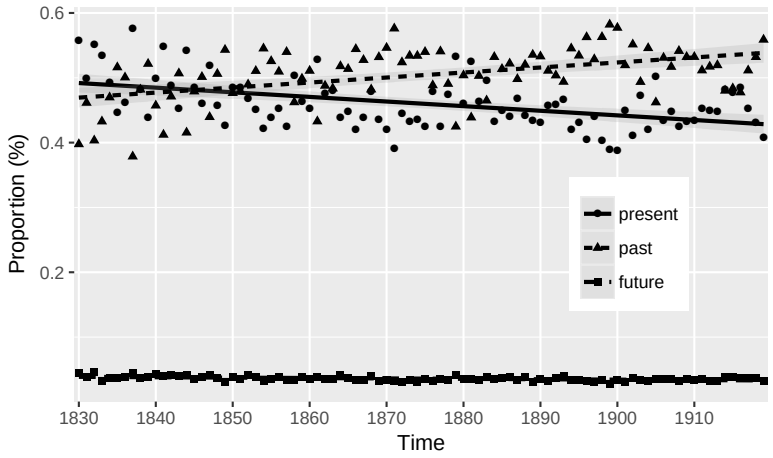
Figure 4.1 – Reference corpus: first-person singular and plural pronouns.

[6]Hereafter also referred to as: '1PL'.

Change in Pronouns Figure 4.1 depicts first-person singular[7] and plural pronouns in the reference corpus over the time span from 1830–1919.[8] As can be observed, 1SG pronouns slightly increase in relative frequency over time. All model parameters in Table 4.3 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 while P&S's Author project has a low non-significant correlation, P&S's Disclosure project shares this highly significant downwards trend. The linear model results mirror the correlation results for both variables. There is less evidence of background language interference in the case of 1SG pronouns, but strong indications for the 1PL pronouns examined here.



(a) Past, present and future tense with respect to relative frequency.



(b) Past, present and future tense with respect to proportion to each other.

[7] Hereafter also referred to as: '1SG'.

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

Change in Tenses Figure 4.2a shows relative frequencies for past, present and future-tense, and Figure 4.2b shows proportions of these three tenses with respect to each other. Present tense appears to vary little over time, and future tense even less, while past tense clearly increases over time. The proportions shown in Figure 4.2b reflect the relative frequency relations in Figure 4.2a. Future tense shows little variation over time or at least not at a significant level, while examining Table 4.3 shows that both P&S's data sets have a positive association for future tense over time. Present tense stays stable in relative frequency, but drops in proportion as past tense increases in relative frequency over time. Present tense has a significant positive quadratic trend as can also be observed in P&S's DP. Past tense in the RC has a highly significant positive correlation (0.7^{***}) and highly significant regression coefficient β, 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^{**}). P&S's 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 similarly to 1PL pronouns could introduce noise into stylistic or ageing analyses, as effects due to general language change might be attributed to either changes in individual style or general properties of ageing.

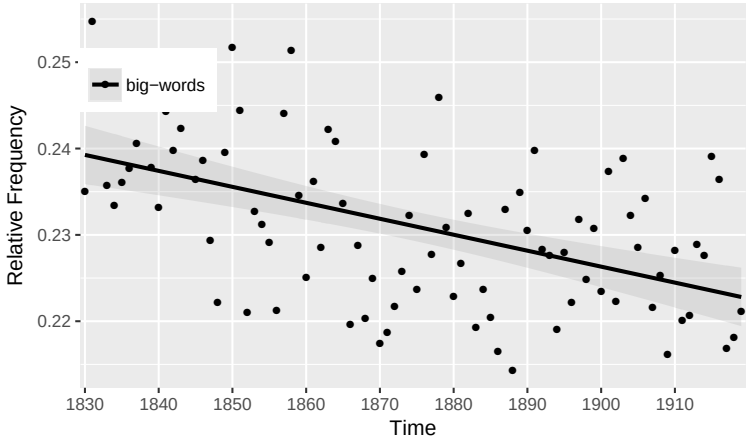


Figure 4.3 – Reference corpus: long-letter sequences.

Change in Big words The development of long-letter sequences over the reference corpus is shown in Figure 4.3. There is a continuous downward trend visible, which is confirmed by both highly significant correlation coefficient ρ (-0.51^{***}) and linear regression coefficient β ($-0.02^{!***}$) in Table 4.3. Both P&S's DP and AP have positive trends and therefore trends in the opposite direction from the reference corpus (r of 0.13^{**} and 0.10 respectively).

Discussion This section has examined six linguistic variables in a continuous section of general language usage, that have been hypothesised in the literature to be affected by ageing in

individual writers. Pennebaker and Stone [2003] have found significant decreases for 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 period. 1PL pronouns were negatively associated for the P&S's Disclosure study and the reference corpus also showed a highly significant negative trend over time. Pennebaker and Stone's [2003] work observed a significant decrease in past tense verbs in the DP. With respect to language change, a general increase in past tense usage was visible and a combined analysis would be necessary to disentangle these effects. Present tense had not been 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, although taking this information into account may still be important. Similarly, there did not appear to be a very strong effect for future tense in the reference language, whereas it was found to increase over all of P&S's data sets, possibly rendering this 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. It is uncertain whether their data was subject to similar background language effects and this cannot be confirmed without examining the relevant language sample appropriate for that time frame. If they were, then this may 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 the original study is not further examined here, as in order to estimate the impact of the background language change, one would need to have access to the original data sets, as well as assemble matching background language corpora. Rather, the remainder of this section addresses how these underlying influences can be taken into account when examining linguistic ageing variables in the literary authors' corpus. Section 4.3.2 tries to estimate the impact of background language change more systematically for the literary authors and considers to what extent, this underlying change influences interpretation of effects previously only attributed to ageing.

4.3.2 Estimating Impact of Language Change

In this section, I aim to investigate the ageing hypotheses with respect to the literary authors corpus *controlling* for background language influence. Analysing different authors together based on their age at the time of publication may introduce several issues for valid inference. The first more general issue is that of background language change, as has been shown here can raise questions as to how much variation in the individual is to be attributed to mere background language change. The second issue resulting from this possible underlying change is that of alignment. Taking the age of an author at time of publication for a set of authors disregards that they were born in different years and not subject to the same influences of background language. The Author project by Pennebaker and Stone [2003] also makes this assumption

as authors' works are sampled from ∼1591–1939. In order to discount possible influences of general language change, a random effects model as shown in Eq (4.5) 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}. Eq (4.6) shows the equivalent quadratic model.

$$y_{ij} = \beta_0 + \beta_1 ref_{ij} + Age_{ij} + A_j + \varepsilon_{ij} \tag{4.5}$$

$$y_{ij} = \beta_0 + \beta_1 ref_{ij} + Age_{ij} + Age_{ij}^2 + A_j + \varepsilon_{ij} \tag{4.6}$$

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 4.4 shows author labels and IDs that are introduced here in the interest of space and referred to in the following discussion.

Table 4.4 – Literary Authors' Abbreviations

ID	Abbreviation	Author
1	ab	Alice Brown
2	amd	Amanda Minnie Douglas
3	cdw	Charles Dudley Warner
4	cfw	Constance Fenimore Woolson
5	es	Edgar Saltus
6	espw	Elizabeth Stuart Phelps Ward
7	ew	Edith Wharton
8	fmc	Francis Marion Crawford
9	ga	Gertrude Atherton
10	haj	Horatio Alger jr
11	hbs	Harriet Beecher Stowe
12	hj	Henry James
13	hm	Harold McGrath
14	lma	Louisa May Alcott
15	mh	Marion Harland
16	mt	Mark Twain
17	rwc	Robert W. Chambers
18	sw	Susan Warner
19	tsa	Timothy Shay Arthur
20	us	Upton Sinclair
21	wdh	William Dean Howells
22	wta	William Taylor Adams

Table 4.5 – This table shows the main model coefficients for simple linear regression using random effects models. 'Age.std' refers to the standardised age predictor and 'Ref.std' to the standardised background change factor. 'Model type' specifies what type of model was used and the last two columns describe the data subset used for that variable, i.e. size of support and ids indicating authors' timelines (compare to Table 4.4). Significance is indicated by: *: $p \leq 0.05$ / **: $p \leq 0.01$ / ***: $p \leq 0.001$ / '$\cdots$' : $p \leq 0.1$. A '†' on the ageing coefficient indicates that the equivalent model using 'year of publication' was more significant.

LIWC Variable	**Model Coefficients**		**Model type**	**Support**	
	Age.std	Ref.std		No.	ID
Social and identity					
First-person singular	0.0008	−0.0002	lme	12	1,4,5,7,10,11,13,14,15,17,19,20
First-person plural	0.02	$0.07^{\cdots}$	glmmPQL:log	15	1,2,3,4,5,6,9,11,13,14,16,17,18,20,22
Time orientation					
Past-tense verbs	0.0008	−0.0002	lme	10	4,8,9,10,13,16,17,18,20,22
Present-tense verbs	0.00009	−0.0004	lme	18	1,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,20
Future-tense verbs	$-0.0002^{***\dagger}$	0.0005	lme	20	1,2,3,4,5,6,7,8,9,11,12,13,15,16,17,18,19,20,21,22
Cognitive complexity					
Big words (> 6 letters)	0.0007	−0.002	glmmPQL:log	21	1,2,3,4,5,6,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22

Table 4.6 – This table shows the main model coefficients for quadratic regression using random effects models. 'Age.std^2' refers to the standardised age predictor and 'Ref.std' to the standardised background change factor. 'Model type' specifies what type of model was used and the last two columns describe the data subset used for that variable, i.e. size of support and ids indicating authors' timelines (as outlined in Table 4.4). Significance is indicated by: *: $p \leq 0.05$ / **: $p \leq 0.01$ / ***: $p \leq 0.001$ / '$\cdots$': $p \leq 0.1$. A '†' on the ageing coefficient indicates that the equivalent model using 'year of publication' was more significant.

LIWC Variable	**Model Coefficients**		**Model type**	**Support**	
	Age.std^2	Ref.std		No.	ID
Social and identity					
First-person singular	-0.00001	-0.5	lme	13	3,7,9,10,11,12,13,14,15,16,17,19,20
First-person plural	0.03	$0.07^{\cdots}$	glmmPQL:log	15	1,2,3,4,5,6,9,11,13,14,16,17,18,20,22
Time orientation					
Past-tense verbs	-0.0002	0.1	lme	10	4,8,9,10,13,16,17,18,20,22
Present-tense verbs	-0.00001	-0.02	lme	18	1,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,20
Future-tense verbs	$-0.000001^{\dagger}$	0.2	glmmPQL:log	18	1,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18,19,20
Cognitive complexity					
Big words (> 6 letters)	0.005	-0.002	glmmPQL:log	21	1,2,3,4,5,6,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22

```
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
```

Figure 4.4 – R output for a glmmPQL-based model predicting future tense from reference language and *age* or *year*.

Table 4.5 shows the results for 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 or log-normal and the final two columns list the subset used for model computation. Overall, there is little evidence for either a very strong influence of background language change or linguistic ageing. The only nearly significant reference language coefficient is 1PL pronouns. Figure 4.4 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. 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 4.6 shows the results for computing quadratic random effect models for the six variables based on Eq (4.6). 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 as in the case of the simple linear model in table 4.5. Although the ageing predictor for future tense in Table 4.6 is not significant, the equivalent quadratic year predictor is.

Finally, I turn to another aspect of this analysis, namely the question of stylistic differences between authors. In order to investigate whether there is likely to be anything particular about Twain's and James' style development with respect to the remaining authors in the set, different aspects can be considered. The previous analyses suggest that the two authors are sufficiently similar to their peers in style to feature in the same models, as evidenced the *Support* column in Table 4.5 and Table 4.6. Further, I consider the specific case of first-person pronouns. Figure 4.5, Figure 4.6, Figure 4.7 and Figure 4.8 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

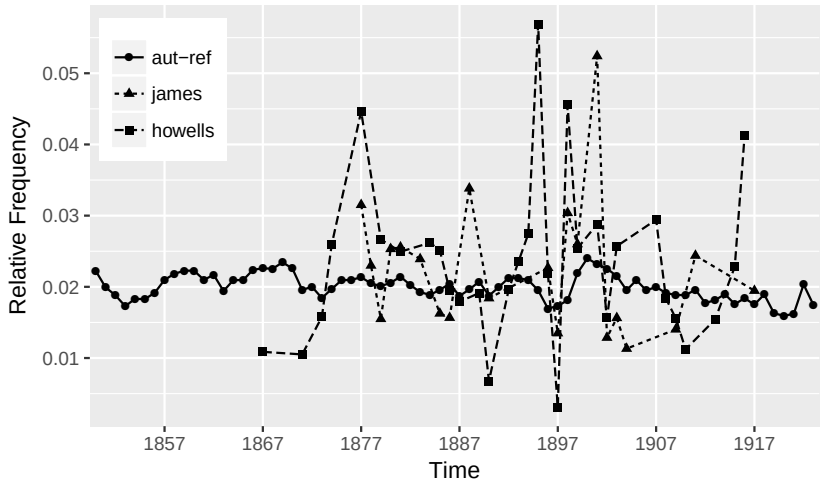


Figure 4.5 – 1SG pronouns for Henry James, William Dean Howells and the author reference corpus (ARC).

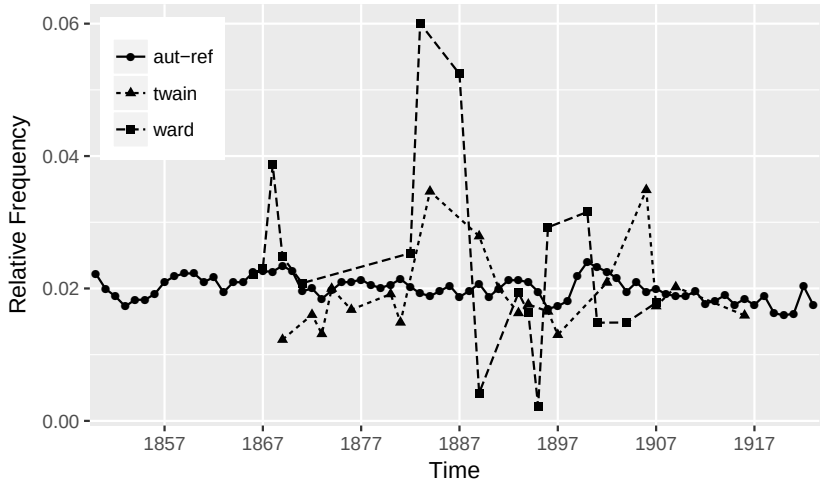


Figure 4.6 – 1SG pronouns for Mark Twain, Elizabeth Stuart Phelps Ward and the author reference corpus (ARC).

over all authors in the set.[9] Figure 4.5 shows Henry James and William Dean Howells and Figure 4.6 shows Mark Twain and Elizabeth Stuart Phelps Ward. For neither Twain nor James, there appears to be a particular development in the form of a trend for 1SG pronouns. Neither 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 4.7 and Figure 4.8 confirm this general impression. For 1PL pronouns, James shows

[9]The 'aut-ref' line represents an average over all authors in the set, computed by: for each year, the average over a feature is computed by 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'.

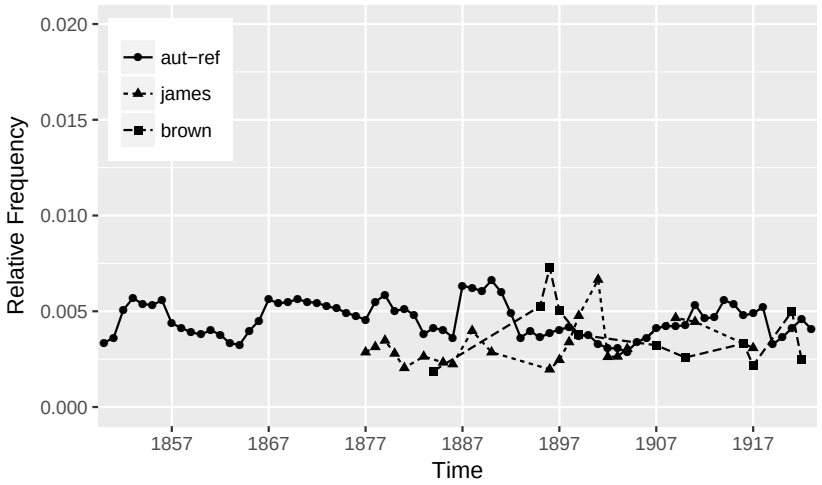


Figure 4.7 – 1PL pronouns for Henry James, Alice Brown and the ARC.

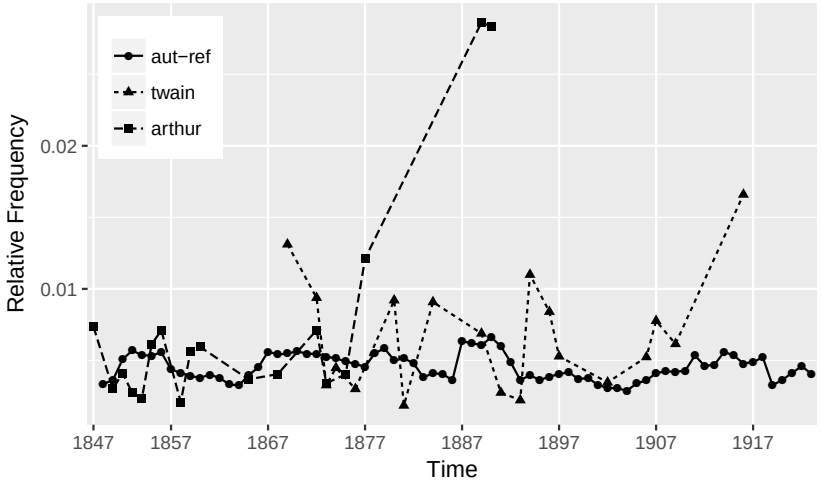


Figure 4.8 – 1PL pronouns for Mark Twain, Timothy Shay Arthur and the ARC.

comparatively little variation over time, while Twain's style displays 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 for Timothy Shay Arthur relative frequency also increases in his last works similar to Twain in his last work. Thus, there appears to be little evidence that Twain and James are decidedly different from their contemporaries in terms of style change.

Overall, there is little evidence that there is a systematic influence of age or background language for the literary authors at least for the variables examined. This could mean that literary authors have a higher command over their language usage and may be more impervious to outside influences.

4.4 Discussion

This chapter 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 P&S 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 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 this linguistic ageing effect possibly not existing for professional writers that could conceivably possess a higher command over their language style than non-professional writers. This would be consistent with P&S findings in so far that their results for literary authors were also less significant than those for non-professional writers. This does not necessarily challenge that linguistic ageing as a property does not exist, but rather suggests that the variables analysed here do not provide good proxy measures at least not with respect to literary writers. However, for this analysis no fully 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 chapter was to examine these six variables for evidence of language change and the results indicate significant change in 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 the 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 Mark Twain and Henry James. The fact that their data was featured in most models suggests that they aligned well their contemporaries on average.

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 are explored in more depth as part of the next chapter looking more specifically at stylistic change in the reference language and the literary authors. Since there do not seem to be stronger differences between James or Twain and the remainder of the literary authors than within the literary authors overall, these are considered as one set after this. As there is no strong evidence of ageing effects in the literary authors, these are not modelled explicitly as part of the following stylistic analysis.

4.5 Conclusion

This chapter 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 previ-

ous 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 in studies like this as has been exemplified as part of this work.

Chapter 5

Elements of Stylistic Change

This chapter specifically considers stylistic change over time for different feature types and ngram lengths with respect to the reference corpus and the literary authors' corpus. Since there was no palpable effect of ageing onto the literary authors' style, this is not considered specifically as part of the subsequent work, not asserting however that there could not be an effect onto the linguistic variables examined here, but emphasising that the focus of this chapter is on stylistic change in literary authors and the interaction between their change and the reference language change. Investigating possible ageing effects on these variables is left for future work; the analytical tools to support this have been introduced in Chapter 4. This chapter shows how to detect salient features for both the reference language and the literary authors and how then to estimate the effect of background language influence on an individual writer.

5.1 Introduction

The last chapter examined specific linguistic variables with respect to how these are influenced as the writer ages, taking into account background language changes at the same time. This chapter considers stylistic language change in individuals comparing to the background language change during the same time period. For this purpose, specifically features that are attested in each time slice' sample of the diachronic corpus are studied and hereafter referred to as 'constant' features. This classification captures occurrence patterns rather than variation in terms of relative frequencies, which may or may not change over the time intervals examined. The method introduced here relies on a temporal prediction task based on the features' relative frequencies to identify salient constant features that exhibit linear change over time. This extends work by Klaussner and Vogel [2015] on predicting the publication year of a text using syntactic word features.[1] That work considered both a data set comprising works of Mark Twain and Henry James as well as a corresponding reference corpus from the $19^{th}/20^{th}$ century,

[1] These are lexical features that have been marked for syntactic function to differentiate between lexical representations that can appear in different syntactic contexts (see section 5.2.1).

sampling features that appeared in many, but not necessarily all time slices. For the two-author data, a root-mean-square-error (RMSE) of 7.2 on unseen data (baseline: 13.2) was achieved; whereas the model built on the larger reference set obtained a RMSE of 4 on unseen data (baseline: 17).[2] Klaussner and Vogel [2018b] extended this work, focusing more specifically on detecting language change for Henry James and Mark Twain in comparison to background language change, only using the prediction task for detecting changing features over time. Four different feature types and ngram sizes were examined, adding character, word stem and syntactic (part-of-speech tag) features to the previous set of only syntactic word features. The results suggested the reference language changed in broader patterns for the constant features, for instance expressions, such as *matter of fact* and *is going to be* increased over the time span of 1860–1920. For James and Twain, analysed based on both their shared and non-shared constant features, different findings emerge: Twain's language features more 'existential' constructions, such as ⟨there's⟩, which James also uses but with less variety. For constructions containing *there*, relative frequency increases in all three corpora over time, although Twain's usage stays consistently higher than both that of James and the reference corpus. For body references, such as e.g. *face*, *eyes*, both Twain and James' average usage lies above the reference corpus for singular items and below it for plural items. Both feature types show a decrease in relative frequency over time. As for stylistic differences between the authors, although they have common features, such as ⟨WDT,⟩, ⟨MD,⟩, and ⟨,., by.IN⟩,[3] actually examining language samples reveals that they use these differently stylistically, where James appears to build very long and intricate sentences that may have contributed to James' later style to be considered somewhat "obscure" and "over-planned" [Beach, 1918]. These three features increase in James' language over time while this effect is present for neither Twain nor the reference corpus.

This chapter continues the work described by Klaussner and Vogel [2018b] in that it considers a larger literary corpus to highlight some aspects of how these authors changed with respect to their shared constant features as a group and compared to the reference corpus. This also provides more interpretative background to the earlier analyses of James and Twain, indicating what aspects they shared with other contemporary authors.

For this analysis, the works of the authors are kept separate, acknowledging their unique development over time. Although it is likely, that individual pairs of authors share aspects of style change, it is unlikely that there is something like a systematic group change (i.e. an 'author tribe') that is not already taken into account as part of the reference language influence. If this did exist, it would require that each and everyone of these authors to have read each other's work or even have collaborated, causing them to consciously or subconsciously repeat

[2]Hereafter, when RMSE is reported, the units are meant to be years the unit is not necessarily repeated. This is to be understood with respect to the caveat that the data is processed using only integer values of years. It is not the case that temporal prediction for any text can be wrong by '7.2 years' - rather by seven years or eight years. The RMSE is an aggregate.

[3]These sequences are constructed based on POS tags as is explained in section 5.2.1

and share each other's language change patterns. Some reference to the average style over the group is given, although interpretation of effects are kept on the basis of individual authors. Although I do not attempt to answer the question of the existence of such a phenomenon, I try to approximate this somewhat by creating an average over the literary authors' corpus, in the following referred to as 'author reference corpus' or 'ARC' that is primarily used for illustrative purposes.[4]

The final step in the analysis is to account for reference language influence in stylistic literary change as has been done before as part of section 4.3.2 wit respect to *age*. Again, the focus is to introduce methods for temporal stylistic analysis, showing some interesting results rather than providing an exhaustive treatment of literary style change. The methods for reasoning about temporal change in constant linguistic features use standard techniques from regression analysis, particularly parameter shrinkage.[5] This research addresses the following question: In what way does literary style change over time with respect to the features that are always there and in what way does it change differently from background language change.

This chapter is structured as follows: Section 5.2 presents feature preprocessing and experiment results. Section 5.3 continues by discussing these and Section 5.4 concludes and summarises the results of this chapter.

5.2 Experiments

Section 5.2.1 introduces feature preprocessing. Section 5.2.2 addresses the general experiment design and model and parameter selection. Section 5.2.3 reports on the results and exemplifies how the resulting models can be used to detect salient features that change over time. Section 5.2.4 considers how to approximate language change influence in the case of an individual author.

5.2.1 Feature Extraction

For the experiments in this chapter, features of different levels of abstraction and ngram length were used. Table 5.1 lists all four feature types ordered by increasing degree of specificity with examples for unigram and trigram size. The most general type is character ngrams, including punctuation and single spaces.[6] While the character ngrams reduce words and sentences to their orthography, the part-of-speech type generalises them as sequences of syntactic types.[7] Word stems present a more specific generalisation of the simple word feature, but rather than

[4] This has previously been introduced as part of Section 4.3.2.

[5] The resulting set of features identified is a specific subset of features that are both constant and have a linear relationship with the response variable over time, i.e. change in trend rather than in periodicity – this is not to say that non-linear patterns or estimation are not interesting, but that for this work the focus is on this particular setting.

[6] Multiple spaces were reduced to single spaces.

[7] To extract part-of-speech features needed for both POS as well as syntactic word features, the TreeTagger POS tagger [Michalke, 2014; Schmid, 1994] was used.

Table 5.1 – Feature types

ngram type	Example	
	unigram	*trigram*
character	⟨c⟩	⟨ca,⟩
part-of-speech (POS)	⟨NP⟩	⟨IN DET NP⟩
word stem	⟨allud⟩	⟨to allud to⟩
syntactic word (lexical)	⟨like.IN⟩	⟨like.VB the.DET others.NNS⟩

capturing syntactic aspects, this type captures what lexical type of word (or sequence) was used, such as ⟨allud to⟩ in place of *allude to* or *alludes to*.[8] The most specific feature is termed 'syntactic word' sequences, meaning words that have been marked for syntactic class, as in the case of *like*, which may be used as a preposition or a verb, depending on context. Compare *I'm like my father.* and *I like my father.*: in the first instance 'like' is used as a preposition, in the second it is used as a verb. Hence, for this feature type, each word is given the correct part-of-speech tag, thus allowing distinct features to be identified for words with more than one syntactic context, such as ⟨like.VB⟩ for verbal usage and ⟨like.IN⟩ for prepositional usage.[9]

Having extracted all feature types, these were then transformed to lowercase, as for this work features are not analysed with respect to sentence boundaries. Finally, document-feature matrices were constructed for each type and ngram size and relativised in the following way: for all of the analyses reported on here, relative frequency relativisation using the total token count was used to account for differences in text available for each year and the counts for each feature.[10] The RC files were simply joined by year. For the LAC, if there was more than one work for a particular year per authorial source available, these were joined together and relativised as one text for each author separately.[11] For both the reference set and the literary authors' set, an ordinal variable *year* was added for each experiment to mark the publication year of a text. Additionally, for the literary authors, categorical variables indicating authorship (A) were also included. The literary authors were then joined into one set creating distinct timelines for authors writing in parallel through the addition of the author variable.

[8] The feature remains orthographic inasmuch as the stem differs from the lemma. Word stems were extracted using the *RTextTools* package [Jurka et al., 2012]

[9] Punctuation and sentence endings are also included as features and in relativisation. The POS tags assigned by the tagger to the individual word entity in its context are used to augment or replace the word entity.

[10] The long and rarer n-gram sequences could cause the data to become rather sparse and feature values thus became computationally expensive. To overcome this challenge, memory-intensive processing steps were separated and simplified relying on both the R packages *bigmemory* [Kane et al., 2013] and *foreach* [Revolution Analytics and Weston, 2014].

[11] Joint relativisation was avoided as it might distort individual differences or create a shift towards authors with more data in a given year.

5.2.2 Model Parameters

For all experiments, the *elastic net* model configurations were used as introduced in Section 2.6, combining both *lasso* and *ridge* regression. Prediction accuracy is evaluated through the root-mean-square-error also introduced in Section 2.6.

To construct the input for each of the 30 models shown in Table 5.3, the same procedure was performed for all of the previously constructed document-feature matrices.[12] The data was first divided into training and test data using a 75/25 stratified split on the ordinal variable *year* that was added in the previous step.[13] After this, all constant features over the training set were extracted, i.e. the features appearing in all training set instances. This subset is then passed to the elastic net models.[14] The final model was computed by performing 10-fold cross-validation on the training data to find the 'best' α and λ parameters, deciding to what extent features were either shrunk or removed from the model as part of the elastic net configuration.[15] The *best* α and λ parameter estimates for a model were set as their combined global optimum. This optimum in turn was defined as the most parsimonious model within 1 standard error (SE) of the model with the lowest error, as defined by the MSE. Not choosing the best performing model allows one to circumvent models that might be needlessly complex and thereby somewhat balancing prediction accuracy and model complexity. The evaluation parameter, RMSE, for the training and internal test set was computed by taking the model MSE and computing its square-root. The above procedure was performed four times, each time selecting a different training-test split and so each model RMSE in Table 5.3 represents the average over four iterations.

Table 5.2 – Baseline for both data sets.

data set	**rmse**	
	training	*test*
literary-authors corpus	18.2	29.0
reference corpus	28.8	29.0

Table 5.2 shows the baseline results for both data sets using the RMSE, i.e. accuracy of only using the intercept through the data set, known as the 'null model'. Linear model assumptions were checked visually by using the *Plotmo* package [Milborrow, 2017].

[12] There were no shared constant stem and syntactic word tetragram for the LAC, hence no models could be computed for these features.

[13] This step was done using the *caret* package in R [Kuhn, 2014].

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

[15] The following procedure was outlined by Nick Sabbe: http://stats.stackexchange.com/questions/17609/cross-validation-with-two-parameters-elastic-net-case – last verified: August 2018

Table 5.3 – Results for the RC (left) and LAC (right) for all four feature types, the first two columns showing RMSE over training and test set and 'model' lists model specifications, i.e. number of coefficients β.

	Reference set			**Literary authors' set**		
type-ngram	**rmse**		**model**	**rmse**		**model**
	training	*test*	βs	*training*	*test*	βs
Char-1	4.3	3.9	14	15.4	12.3	18
Char-2	4.3	4.1	65	11	9.7	133
Char-3	3.8	4.0	310	8.7	8.6	371
Char-4	3.4	3.5	546	8.3	8.3	322
	training	*test*	βs	*training*	*test*	βs
POS-1	5.3	5.6	14	13.6	11.6	21
POS-2	4.1	3.9	94	9.9	9.4	144
POS-3	3.9	3.6	963	10.1	9.7	180
POS-4	4.1	3.9	3278	11.2	11.5	101
	training	*test*	βs	*training*	*test*	βs
Stem-1	3.4	3.7	121	10.2	10.1	94
Stem-2	3.6	3.8	125	17.5	20.3	14
Stem-3	4.0	4.7	146	17.8	20.6	4
Stem-4	5.5	5.5	164	NA	NA	NA
	training	*test*	βs	*training*	*test*	βs
Lex-1	3.5	4.1	148	9.4	9.5	118
Lex-2	3.6	3.8	487	17.5	20.3	21
Lex-3	3.8	4.3	203	17.7	20.6	10
Lex-4	4.8	4.6	983	NA	NA	NA

5.2.3 Literary Style Change

Table 5.3 shows the results for the linear regression analysis for both reference corpus and literary authors' corpus. Elastic net models tend to select groups of correlated predictors rather than only one of a group of related predictors, explaining why model size generally grows with ngram window size. A previous analysis by Klaussner and Vogel [2018b] has shown that if sparser models are desired, the corresponding *lasso* model can be used without great loss of prediction accuracy. For the purpose of detecting salient features, the present configuration is more useful as it allows detection of features that are somewhat related in their change patterns. Overall, the results for the reference language corpus are very regular, varying between 3.6 and 5.5 in error for both training and test set and thus being far below the baseline of 29 years deviation for those sets if no external predictors are taken into account. Model size grows much larger for the less specific types of character and POS ngram than for stem or syntactic word ngrams.

For this analysis, I focus on literary style change with respect to background language change and features primarily relevant for only background language change are not investigated further here, but rather salient features of the LAC that also appear in the RC models are considered in more detail. Comparisons between authors are based on randomly extracted samples and are not exhaustive, but rather anecdotal. The previous model results for the literary authors indicated that there is probably considerable variation between them, as less specific types such as character ngrams or syntactic word unigrams tend to return more accurate models. In order to select temporally changing features, ideally one takes those model features that are shared over all four test/training set divisions for a particular feature type and ngram size. Those features with the highest accumulative weight can then be considered more closely. This has the effect that any features that vary with respect to the sign of their weight over the models, averages to a much lower overall weight. The character trigram and tetragram models achieve the highest accuracy overall. Over all iterations, 205 features are shared, where among the highest weighted ones are ⟨od_a⟩, ⟨e_up⟩, ⟨n_fo⟩ and ⟨sed_,⟩. Except for ⟨n_fo⟩, all of them have a negative coefficient indicating that these might have been decreasing in usage for the literary authors.

Figure 5.1 and Figure 5.2 show two character tetragrams for two of the authors: Amanda Douglas and William Dean Howells and the reference set as well as the author reference corpus (ARC).[16] The feature ⟨beca⟩, in the top plot is shared by the RC and LAC, but would not be ranked highest for either of the sets. This character tetragram would be mostly realised by *because* (78%) and some cases of *became* (22%). In comparison, the plot just below shows a more salient feature for the authors that was not part of the features shared with the RC: ⟨n_fo⟩.

[16]The 'aut-ref' line represents an average over all authors in the set, computed by: for each year, the average over a feature is computed by 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.

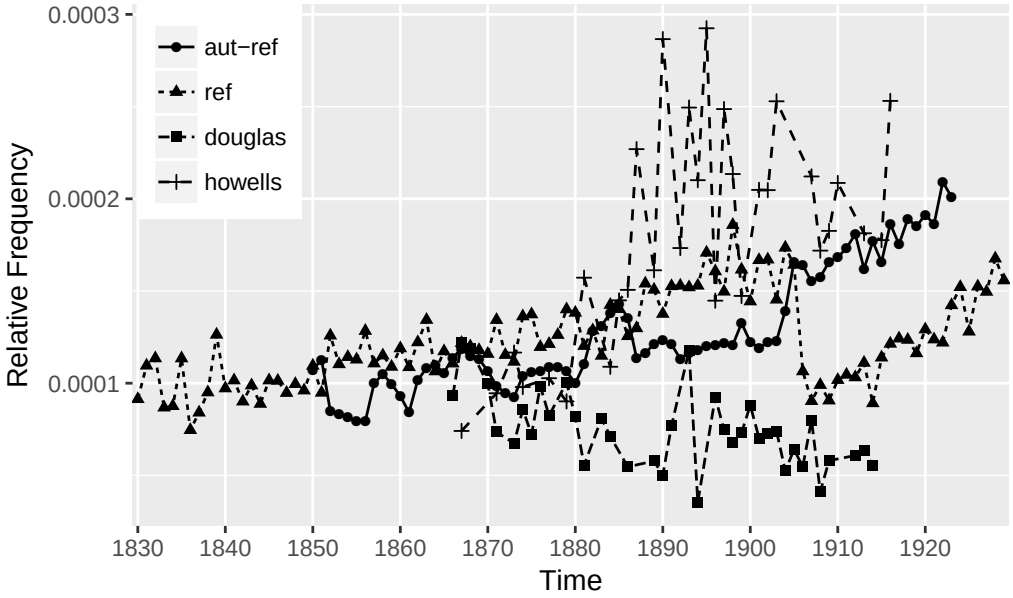


Figure 5.1 – The feature ⟨beca⟩ for Douglas and Howells alongside the RC and ARC.

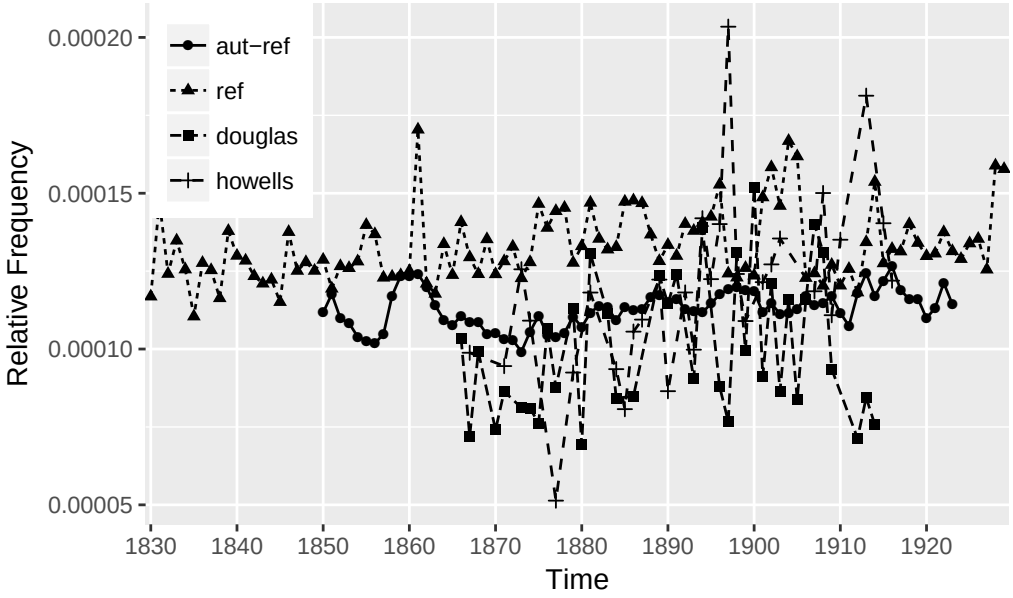


Figure 5.2 – The feature ⟨n_fo⟩ for Douglas and Howells alongside the RC and ARC.

This feature would occur predominantly in constructions with *n for* (64%) as in *If it had not been for you*. The more general sequence n for appears in sentences, such as *She had been fond of him as a child*, subsumes the previous one and accounts for another 20% of occurrences. There appears to be a clearer upwards trend for this feature. Of the part-of-speech ngrams, bigrams are most accurate. Examining the list of 106 shared author model features, shows that ⟨RB RBR⟩, ⟨NN WP⟩, ⟨VBP NNS⟩ and ⟨JJR IN⟩ are the highest rated features, of which

the first three are positively associated and the last one negatively associated. Figure 5.3 and Figure 5.4 show two of these features for the authors Gertrude Atherton and Horatio Alger alongside both the reference corpus and author reference corpus. Both features are not among the highest ranked reference corpus' features, although both corpora show a very similar trend for ⟨NN WP⟩.

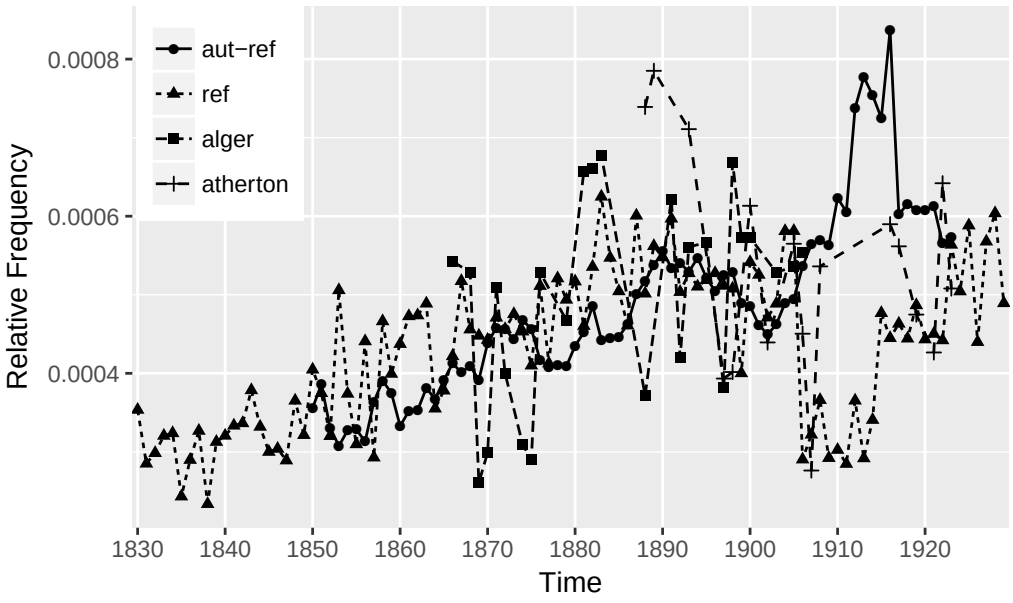


Figure 5.3 – The feature ⟨NN WP⟩ for Alger and Atherton alongside the RC and ARC.

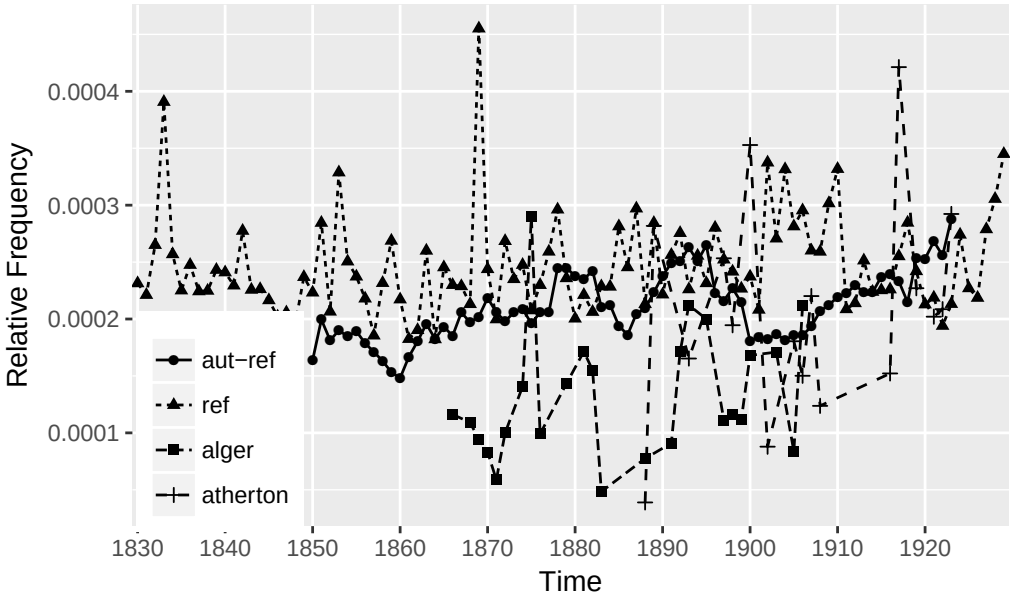


Figure 5.4 – The feature ⟨VBP NNS⟩ for Alger and Atherton alongside the RC and ARC.

The study by Klaussner and Vogel [2018b] reported on a few salient POS bigram features

for Twain and James, among these ⟨MD ,⟩. Figure 5.5 and Figure 5.6 show this feature for Timothy Arthur and Robert Chambers and James and Twain separately. For this feature, Arthur and Chambers appear to be following the general trend of the reference language, whereas James' and Twain's development follow a somewhat opposing trend.

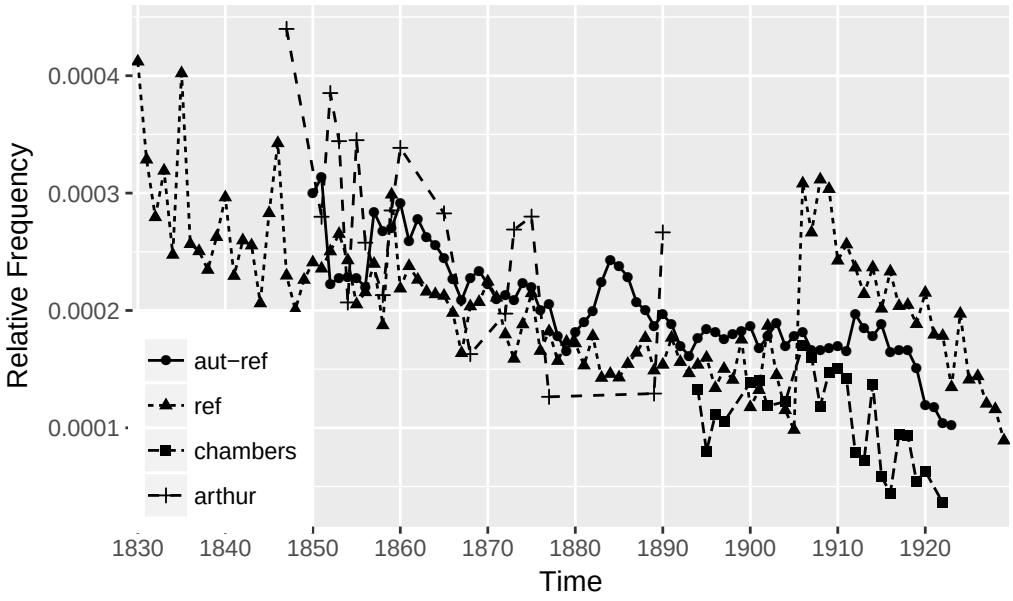


Figure 5.5 – The feature ⟨MD ,⟩ for Chambers and Arthur alongside the RC and ARC.

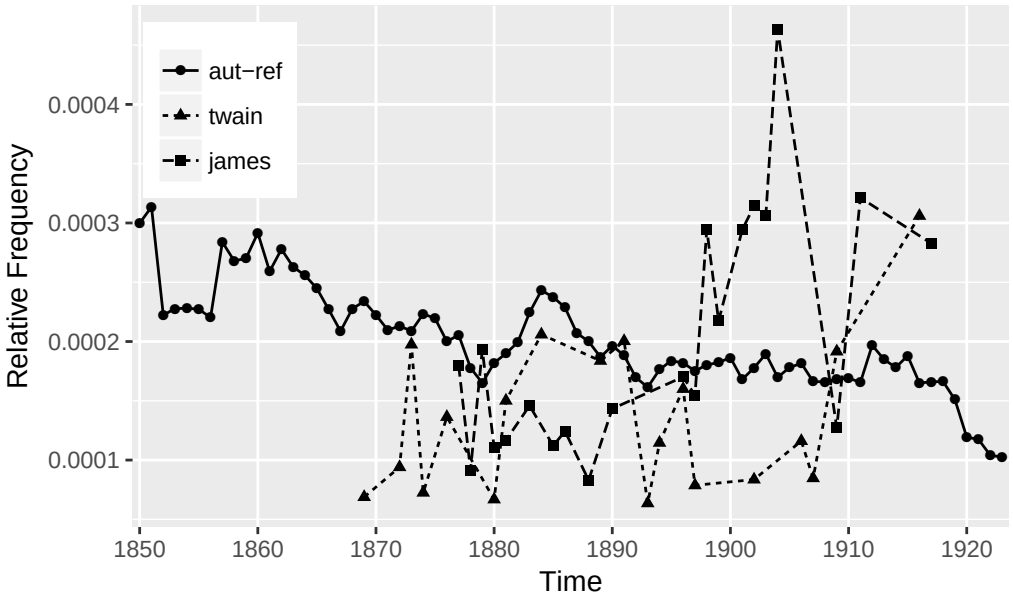


Figure 5.6 – The feature ⟨MD ,⟩ for Twain and James alongside the RC and ARC.

As for salient stem unigrams, ⟨near⟩, ⟨onc⟩, ⟨back⟩ ⟨end⟩ appear at the top of the list, where except for *near* all other items have a positive association over time. Figure 5.7 and

Figure 5.8 show ⟨near⟩ and ⟨back⟩ for Francis Crawford and Elizabeth Stuart Phelps Ward. While only Ward's usage for *near* appears to be lower than Crawford and both the RC and ARC, for *back*, the two often appear to have a higher usage than was representative for that time. The literary authors tend to agree stronger on the features that are ranked higher in their combined models and also show more agreement for features based on models that were more accurate. The models can also be used for individual stylistic analysis; interpretation in that case is based on individual stylistic changes rather than a group of literary writers as was done here. This analysis has shown that the method is valid for detecting linearly changing features over time with respect to literary change.

Style comparison Klaussner and Vogel [2018b] reported on language examples to exemplify how James and Twain use the same features quite differently. They identified two POS bigrams and one syntactic word bigram (⟨WDT,⟩, ⟨MD,⟩ and ⟨,., by.IN⟩), which increase in James' language specifically and where Twain's usage mostly lies below the one for the reference corpus. Several language examples for these were extracted from the two authors' data, e.g. an example of ⟨MD,⟩ for James is shown in (1) and (2) and for Twain in (3), where James uses these features to create very long-winded sentences seemingly not characteristic of Twain's style. Similarly, one can extract sentences containing ⟨MD ,⟩ for the LAC in general. Example (4) shows one of William Howells' sentences containing the same feature. Similar to James, he seems to frequently compose very long sentences, although they appear less intricate, whereas Amanda Minnie Douglas' usage in Example (5) resembles more that of Twain. Example (6) shows a sentence randomly extracted from one of Constance Fenimore Woolson's works, which may be somewhat reminiscent of James' more complicated style, something that could be anchored in the fact that Woolson was an admirer of James' work.

(1) *It sounds, no doubt, too penetrating, but it was by no means all through Sir Claude's betrayals that Maisie was able to piece together the beauty of the special influence through which, for such stretches of time, he had refined upon propriety by keeping so far as as possible his sentimental interests distinct.*

(2) *Then it is, in the final situation, that we get, by a backward reference or action, the real logic and process of the ambassador's view of how it has seemed best to take the thing, and what it...*

(3) *People in giving to me without compensation a book which, as history had afterward shown, was worth a fortune.*

(4) *He did make room for me in his own department for as long as he could, or as I would stay, when I went down to Cincinnati to look the ground over, and he kept me his guest as far as sharing his room with me in the building where we worked together, and where I used to grope my way toward midnight up a stairway entirely black to his door.*

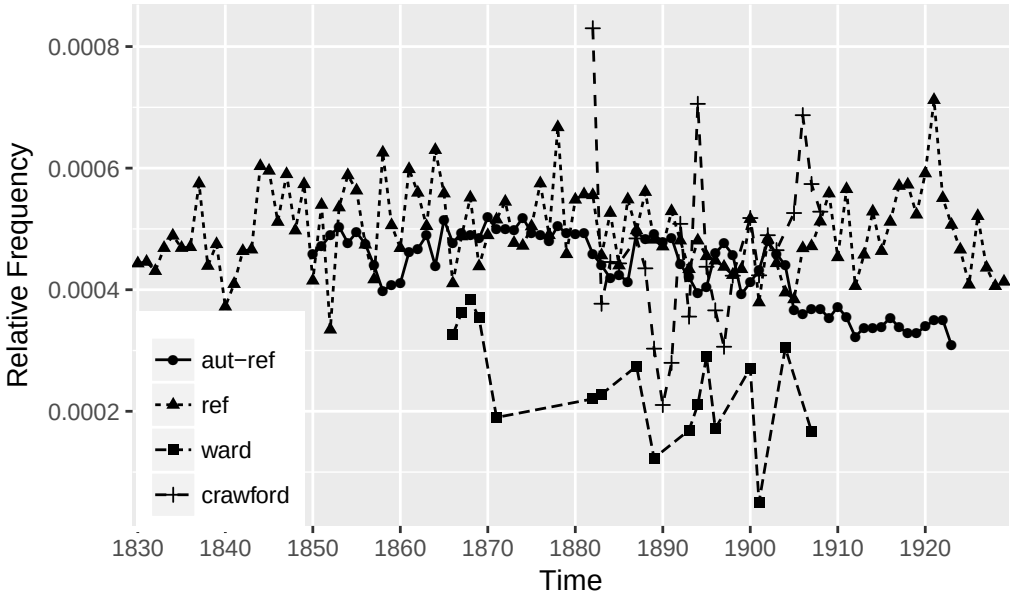


Figure 5.7 – The feature ⟨near⟩ for Ward and Crawford alongside the RC and ARC.

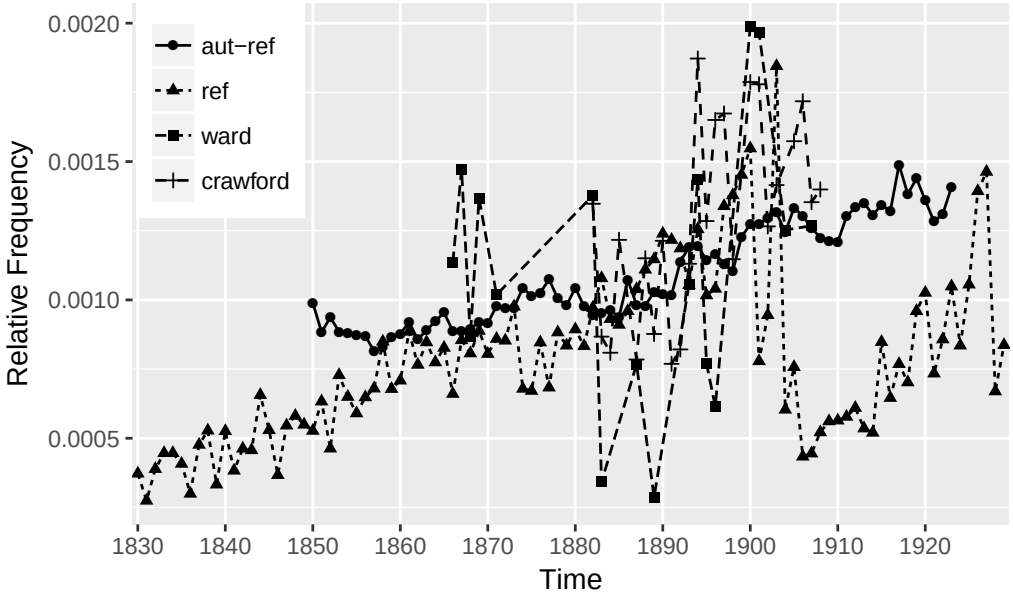


Figure 5.8 – The feature ⟨back⟩ for Ward and Crawford alongside the RC and ARC.

(5) *She would not tell his father just now, but if he ever struck or pinched the babies again she certainly would, and he would be punished twice over.*

(6) *It must, however, be added that the museum will not make this impression upon persons who are indifferent to the general aspect of an aisle, or of a series of walls – persons who care only for the articles which adorn them –the lovers of detail, in short.*

The examples presented here are anecdotal and even though based on randomly extracted sentences would need more investigation and data samples to allow for a more substantiated comparison between the ways authors use stylistic features differently. It served to exemplify, however, how statistical models can be used to identify changing features that could be used for comparisons between styles. This application could be extended to observe possible changes in the way authors use features in context.

5.2.4 Estimating Language Change Influence

Section 4.3.2 has provided a method to estimate the impact of language change on ageing variables. The same can be done for other variables that possibly change stylistically over time. As before, impact through general language change can be estimated for a group of authors by fitting a random effects (linear) model in Eq (5.1), where y_{ij}, the relative frequency of y for author j in year i is predicted using the relative frequency of the reference language in the same year ref_{ij} and the ordinal year itself Y_{ij}, and random error ε_{ij}.

$$y_{ij} = \beta_0 + \beta_1 ref_{ij} + Y_{ij} + Aj + \varepsilon_{ij} \tag{5.1}$$

In this case, stylistic impact for the individual author may be of more interest and Eq (5.1) is simplified to Eq (5.2) discarding the author variable. Another possibility is to simply deduct the reference corpus frequency for that particular variable from the one of the author, however this would naturally assume a linear relationship and would therefore not be adaptable for the the general case.

$$y_i = \beta_0 + \beta_1 ref_i + Y_i + \varepsilon_i \tag{5.2}$$

Figure 5.9, Figure 5.10, and Figure 5.11, show examples of applying this where William Taylor Adams' curve is estimated to rely very strongly on publication year, Gertrude Atherton's curve relies somewhat on both *year* and reference language and Edgar Saltus appears to vary more randomly for $\langle \text{MD} , \rangle$. Figure 5.12 depicts the feature for these three categories of author interaction with reference language.

```
lm(formula = MD.wta ~ ref + YEAR.std, data = dat)

Residuals:
         Min           1Q       Median           3Q          Max
-0.000130359 -0.000047048  0.000006507  0.000035369  0.000124147

Coefficients:
               Estimate   Std. Error t value Pr(>|t|)
(Intercept)  0.000155054  0.000146026   1.062  0.29708
ref          0.721063888  0.582126479   1.239  0.22540
YEAR.std    -0.000005239  0.000001831  -2.861  0.00776 **
```

Figure 5.9 – William Taylor Adams: R output for predicting the relative frequency for the feature ⟨MD ,⟩ from reference corpus frequency and publication year.

```
Call:
lm(formula = MD.ga ~ ref + YEAR.std, data = dat)

Residuals:
         Min           1Q       Median           3Q          Max
-0.000095204 -0.000036394 -0.000005504  0.000020795  0.000105594

Coefficients:
               Estimate   Std. Error t value Pr(>|t|)
(Intercept)  0.000402360  0.000132452   3.038  0.00886 **
ref         -1.371485880  0.714034778  -1.921  0.07537 .
YEAR.std    -0.000006123  0.000003064  -1.999  0.06545 .
```

Figure 5.10 – Gertrude Atherton: R output for predicting the relative frequency for the feature ⟨MD ,⟩ from reference corpus frequency and publication year.

```
lm(formula = MD.es ~ ref + YEAR.std, data = dat)

Residuals:
         Min           1Q       Median           3Q          Max
-0.000211633 -0.000087257 -0.000009018  0.000096950  0.000255192

Coefficients:
              Estimate  Std. Error t value Pr(>|t|)
(Intercept)  0.00058175  0.00066850   0.870    0.401
ref         -1.45400269  3.20911631  -0.453    0.659
YEAR.std    -0.00001352  0.00001589  -0.850    0.412
```

Figure 5.11 – Edgar Saltus: R output predicting the relative frequency for the feature ⟨MD ,⟩ from reference corpus frequency and publication year.

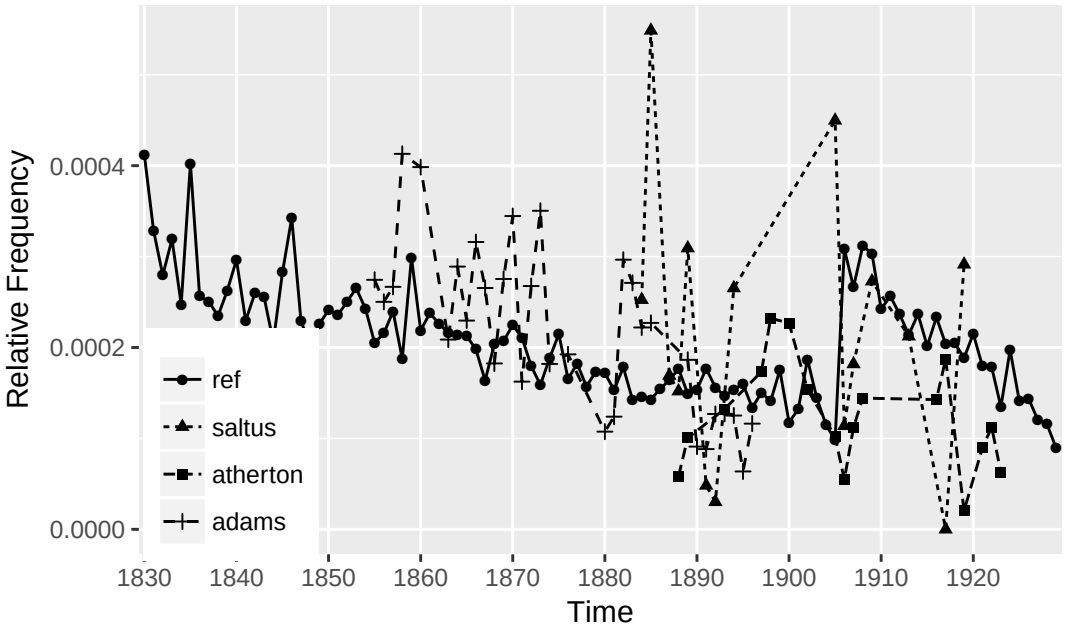


Figure 5.12 – The $\langle \text{MD} , \rangle$ for William Taylor Adams, Edgar Saltus and Gertrude Atherton, alongside the RC.

5.3 Discussion

This chapter has considered the analysis of more general aspects of literary style change and to what extent this change is also existent in the corresponding background language change. For this, four different feature types and sequence sizes were examined. The resulting models return accurate results for both the reference corpus and literary authors' corpus, indicating that for both sets there exist features that change in a linear fashion over time. What is notable is that the more specific feature types, such as stem and syntactic word features are more accurate for unigram sizes whereas character features only become more accurate with increasing ngram size and POS ngrams stay relatively consistent for all four sizes. This could indicate that there is more variation between authors for longer expressions and aspects of style that are more consciously chosen and reflected through stem and syntactic word features, whereas their shared language variety may cause them to use similar syntactic constructions. There also appears to be individual variation in authors that could be explored in more detail. Also similarities between pairs or groups of authors could be considered more specifically to provide insights into shared style change. The question of a group effect, i.e. the existence of common stylistic change between all authors of the corpus caused by mutual influences could not be answered sufficiently as part of this work and is left for future work. Methods for detection of general stylistic features have been proposed and it has been shown that these return accurate results, thus allowing to interpret chosen model features with respect to stylistic differences and language change. Further, a combined model has been proposed to estimate background

language influence for particular variables for individual authors thus somewhat disentangling what is general and what is individual with respect to stylistic change for an author. The methods presented here are equally applicable to smaller sets of literary authors.

5.4 Conclusion

This chapter has presented and exemplified methods for the detection and analysis of change in stylistic features with respect to background language change. The results obtained are accurate thus validating the group of features used for prediction. Further, this research has introduced models to account and quantify underlying language change in the individual author's case to disentangle individual from general effects.

Chapter 6

Interpretation of Stylistic Change

The last two chapters considered influences of ageing and general stylistic changes, where both analyses focused on features that appeared in all time slices of the data period examined. One of the questions that has yet remained unexplored is how constant features relate and are possibly influenced by changes in non-constant features. This chapter continues the analysis of change in features that are attested in all time slices, specifically with respect to change that is more sudden rather than the more gradual change in trend considered as part of Chapter 5. Of particular interest in this is whether other non-constant clusters of words could be related to these sudden frequency changes in constant features and if so which types would be most prevalent.

6.1 Introduction

Change in linguistic variables can occur in different shapes and forms: slow gradual change as opposed to sudden and abrupt, short as well as long-term effects. Differences could be rooted in levels of linguistic abstraction, as for instance individual words are likely to show more variation over time than entire word classes, where smaller fluctuations would be averaged and only larger trends pervading the entire group would be more easily discernible. This chapter considers the analysis of the more regular and possibly also frequent items, in particular those appearing in all years of the time period examined. Typically, these features are more general in meaning (e.g. temporal expressions) rendering them suitable for a variety of language contexts opposed to strongly topic-related words, such as *hurricane* or *computer*. The main hypothesis investigated here is that these items change through other less constant items that are more prone to topic change over time, such as concepts relating to the outbreak of a war or natural disaster. The type of change sought is a change in mean, whereby a feature changes its relative frequency fairly abruptly at time t, rising or falling to a new level and remaining there for at least a time span of 5–10 years. If one compares the mean over the samples before time t to the mean taken over the samples after time t, one obtains significantly different means.

This type of research has to be distinguished from two related areas of research, i.e. collocation analysis and semantic change in the form of 'neologisms'. Semantic change analysis is different in that it considers cases whereby a word acquires a new sense and possibly also a second part-of-speech class and could subsequently be used in different syntactic contexts, whereas here the focus is on changes in word frequencies and their possible non-semantic change related causes, as for instance temporal expressions used together for contrast and thereby often occurring together, e.g. *If I had only known then, what I know now.* Also, conceptually, these regular and irregular appearing words could be relatable through collocations or otherwise longer ngram sequences. While the method presented here could be used for their detection as well, it is not limited to relationships between words that occur close to each other, but also words or expressions that only share a conceptual rather than spatial relationship. For this work, temporal expressions are going to be analysed with respect to the larger reference corpus and correspondingly also for the literary authors' corpus.

Constancy in occurrence in this context is seen as a continuum, ranging from *hapax legomena* or the words that only occur once over the entire temporally-ordered corpus, over features appearing in a few years, to the features that appear in all time slices of the time-period examined. In order to broadly differentiate between these, the constant features are referred to as 'universally constant' and the features appearing in more than 10 percent of the time span as 'partially constant'.[1] For the present analysis, the relationship between partially constant words and constant words is explored. In particular, this analysis tries to relate the change in frequency of words that are *always* there to words that emerge around certain points in time and only remain frequent for shorter periods of time, suggesting they are more prone to changes in popular topics. In spite of mainly using statistical change-point analysis for identification of significant change over time, other means of interpretation and further validation of findings are sought, for instance through examining actual sentences in the data. The methods presented are equally valid for application to other types of word classes. Temporal expressions have the advantage that they are likely to be less context-dependent than for instance verbs (e.g. when analysing the change in use of auxiliary plus contractions, where the surrounding verb types are likely to be important). These methods were first introduced by Klaussner et al. [2017]. The purpose of this study is to see whether sudden change in relative frequency in constant features relates to more changeable words or expressions co-occurring with them and if this can be detected automatically.

The remainder of this chapter is organised as follows: Section 6.2 presents the methods employed. Section 6.3 describes the experiments for reference and literary authors' corpus. Section 6.4 reviews and discusses the results and Section 6.5 concludes this chapter.

[1] By 'universally' the span of our entire data set is meant rather than any data and time space that could be examined in this way.

6.2 Methods

Section 6.2.1 describes methods used for initial detection of interesting constant features, followed by the actual change-point method in Section 6.2.2.

6.2.1 Detecting Changing Features

Words belonging to the same word class can be subject to the same type of change. Hamilton et al. [2016], for instance, found that verbs are more likely to undergo global semantic shifts, whereas nouns are more prone to local or cultural shifts, e.g. emergence of *virus* as in *computer virus*. Although my analysis does not examine semantic change, it is reasonable to assume that word class and frequency level are also important for non-semantic change. The experiments described in section 6.3 are aimed at detecting groups of partially constant words whose occurrence patterns are likely to have changed based on some shifts in popularity of particular topics and which in turn may then influence relative frequency shifts in universally constant features. The focus for this analysis is on cultural or event-based shifts, suggesting the analysis of *noun-noun* and *adjective-noun* bigram sequences for the reason that these would be the common form of representing events, both belonging to open-class categories and likely to follow a similar frequency distribution.[2] Other noun phrase types are discarded as one would expect those other types, i.e. containing determiners, proper nouns and pronouns to have different frequency levels that might introduce noise into the analysis. Bigram rather than unigram size was chosen as it provides more context and is richer in meaning allowing one to discern more specific items of change than with unigram size. Analysing items of higher abstraction, such as part-of-speech sequences could as be more difficult to evaluate, while word sequences offer more possibilities for human evaluation.

In order to differentiate between lexical representations with more than one meaning, word sequences were annotated for syntactic context using part-of-speech tags from the TreeTagger POS tagger [Michalke, 2014; Schmid, 1994], e.g. the word *like* has different meanings depending on its context. It can be used as both a verb and a preposition, which should subsequently be treated as two separate items. The new syntactic word features were then created by using the tag sequence as a suffix to the original word in context that gave rise to it. Thus, *He likes her* becomes *he.PP likes.VBZ her.PP*.[3] Items are then joined to bigram sequences and each two syntactic word sequence is relativised by the total number of bigram sequences in that year. Having collected all noun-noun and adjective-noun bigrams, these are then reduced to the universally constant features. In order to discover interesting (and possibly related) features more easily, this set of universally constant features is ordered according to mean relative frequency

[2]Thus, relationships between the same type of features are considered, rather than between different types of features. This is not to suggest that those combinations would not be interesting to explore, but that this might require adjustment of the current method.

[3]In this, the difference between the original word in context and the lemma of the word would primarily be reflected in verbs.

and then subjected to principal component analysis (PCA) on sets of 50 bigram features ordered by relative frequency, as previous experiments have shown estimation and later interpretation of components to be better, when the document-feature ratio is in favour of more samples.[4]

6.2.2 Change-point Detection

Change-point analysis is the analysis of a time-series with the aim to detect specific points t in time that separate the points before and after it with respect to some criterion. More formally, aspects of change-point analysis can be defined as follows: given a sample of time-series $\{y_t : t \in 1,...n\}$, a change-point occurs if there exists a time k, where $1 \leq k \leq n-1$, such that the distributions of $\{y_1...y_k\}$ and $\{y_{k+1}...y_n\}$ are different with respect to some criterion, i.e. change in *mean*, change in *regression* or change in *variance*. For this analysis, changes in mean are of primary interest as these would signal a different usage of a feature with respect to an earlier time period, whereas changes in variance indicate that the author has become more or less consistent for the feature. The time period should comprise at least 10 years or so, requiring a change-point detection technique that is less volatile to short-term fluctuations in the data. For the experiments here, the approach by James et al. [2016] was chosen, originally used for breakout detection in cloud data in the presence of anomalies.[5] The proposed approach ('E-divisive with Medians'(EDM)) is a non-parametric technique using medians and estimating the statistical significance of a change-point through a permutation test. This technique appears to return sparser, more intuitive results than distribution based change-point methods, rendering it even more desirable in this case as data is not always normally-distributed.

6.3 Experiments

This section specifically considers change in various temporal expressions and possible relations and influences from word clusters that are appearing and disappearing over time. Section 6.3.1 examines change in temporal expressions in the reference corpus' news genre and Section 6.3.2 considers how these expressions change in the literary authors' corpus.

6.3.1 Temporal Expressions in News Data

Some previous data exploration of the reference corpus indicated that its genre types may considerably differ with respect to changes and when these occur and that the news corpus might exhibit interesting patterns of sudden changes with respect to noun-noun and adjective-noun types.[6] This intuition is supported by the results of performing a preliminary feature

[4] PCA is an unsupervised statistical technique to convert a set of possibly related variables to a new uncorrelated representation or principal components. This type of analysis groups features according to common variance patterns and can help to detect features that vary in a similar way.

[5] This method was implemented in the R package *ecp* [James and Matteson, 2013].

[6] Article titles were not included in the source texts.

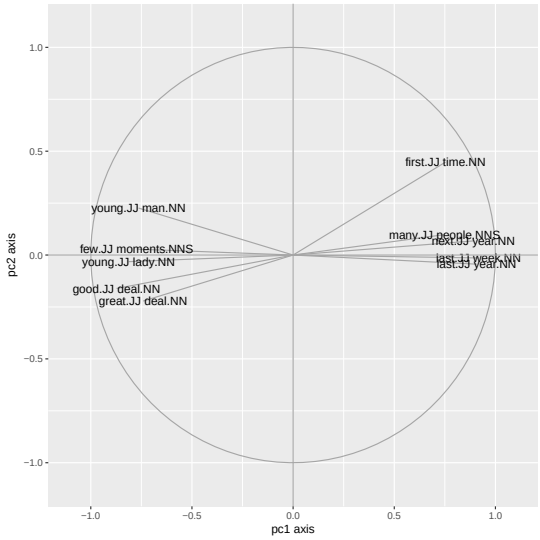


Figure 6.1 – RC: PCA results for the 10 highest associated bigrams.

search as described in 6.2.1, whereby PCA is applied to sets of 50 features ordered by relative frequency.[7] The results of running PCA on the 50 most frequent noun-noun/adjective-noun sequences are 50 new components that group related features together. A feature can be negatively or positively related to a new component. The components themselves account for decreasing proportions of variance, e.g. in this case the first component accounts for 25% and the second component for 12% of the variance with the rest being more broadly spread out. Inspection of first principal component allows for discovery of the three highest associated items: 'last week', 'last year' and 'next year' with similar weights: 0.2596, 0.2569 and 0.2523 respectively as shown in Figure 6.1. Figure 6.2 shows two of the features: *last year* and *last week*. Both display rather sudden changes around 1920, where *last year* becomes primarily more frequent in the news domain and after that magazines, while for *last week*, the order is reversed: first magazine data and then news data. In the following analysis, these expressions are more specifically considered as part of the news section and compared to the other genres when appropriate. For this, an extract of a 100 years from 1880–1979 was chosen covering a substantial time span and most importantly the years from ca. 1920–1930 that appear to have given rise to particularly pronounced changes.[8] For relativisation of feature frequency, all individual files for each year were combined and relativised by the overall token count for that year.[9]

[7] As the data was not normally distributed, logarithms of relative frequencies were calculated before applying PCA.

[8] Some of the years before that, e.g. 1867 have only a few samples and would therefore not be representative.

[9] In the case of higher sequence features, such as word bigrams the unigram token count is replaced by the unique bigram token count.

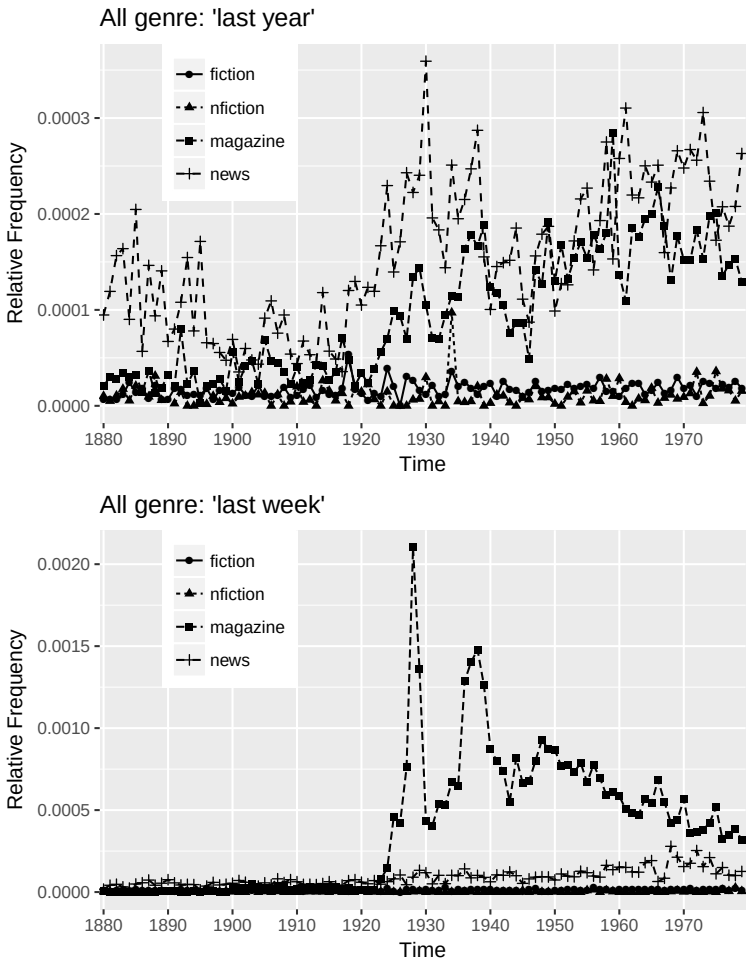


Figure 6.2 – RC: word bigrams *last year* and *last week* shown over different genre types: news, magazines, fiction and non-fiction.

6.3.1.1 Discovery of Related Variables

Given the universally constant feature of interest, in this case a temporal expression, one seeks to find features that display similar changes or as hypothesised are somewhat responsible for the change observed in that universally constant feature. The set of suitable candidate variables comprises the set of bigram adjective-noun/noun-noun combinations, that in contrast to the main variable need not be universally constant, but might only turn out to be partially constant over the entire time span.

Choosing the expression *last year* for the first trial of finding related non-constant features, the very first step in this is to run a change-point analysis for *last year* over the entire 100-year span in order to ascertain the exact point of change. As could be observed from Figure 6.2, a change happened a little after 1920 with the period afterwards giving rise to a higher frequency pattern than the years leading up to it. Having found a change-point for the chosen feature, one then selects an interval of a certain length (e.g. 10 years) after the change-point to limit the number of candidate features for examination. The rationale in this case being that given a rise

in frequency after the change-point one would expect related features to be partially constant for at least a certain period of time afterwards. One therefore extracts the partially constant features for this time period only. One then takes the remaining features and calculates their individual change-points over the entire 100-year period. Given all change-points over all features, these are divided into two different groups of features, those whose change-points occur before or exactly at the same time as the main feature's change-point and those whose change-points occur after. Only those features that change significantly with respect to their mean within 10 years before or after the main feature's change-point are retained, the reason being that it is deemed less likely that those changes more remote in time would be related. Using the present method for detection, features usually do not have more than one change-point and in the cases that they have two, these are usually separated by a time span of at least 20 years. A change-point in the present setting indicates a change in mean and what follows could either be an increase or a decrease in frequency.[10]

As this exploratory work only focuses on features with similar trends, those features with opposing trends to the candidate feature are discarded. This decision is based on calculating the correlation between *last year* and each feature over the interval covering 15 years on either side of a feature's change-point and only retaining those features for which this correlation is positive.[11] In the present case, the specifications were set as follows: the change-point for *last year* was estimated at 1923, so I chose the interval spanning the years 1924–1934 to look for features that are constant over this period of time. One would not expect the exact time frame to be of high importance, as most features probably level off more gradually over time. After discarding features not constant over this interval, 103 features are left, where at least 22 of these are also temporal expressions. In fact, when one considers the universally constant adjective-noun combinations that are constant over the entire 100-year news data span, the majority of these turn out to be temporal expressions (12 of 16 universally constant expressions). The fact that not more features are constant over the entire span hints at the domain being somewhat volatile with respect to content type sequences.

Table 6.1 shows the highest pairwise correlations (either negative or positive) between *last year* and each of the 103 features over smaller intervals of 10 years from 1920 to 1970, where the universally constant features are marked in italics. The first interval covers a few years before the change-point and a few years after that, so somewhat of a transition period where different concepts have similar trends to *last year*. There are a few temporal expressions and politically/industry-related terms, such *floor leader*, *executive session*, *vice president* and *automobile industry* and a few expressions (possibly temporal), that would probably be anchored more strongly in the business context, such as *first quarter* and *second quarter*.

[10] The focus here is on parallel changes and causes, i.e. the parallel increase of two features together, rather than assuming that a decrease in one feature causes an increase in the other feature, although this would also be a valid scenario.

[11] For this evaluation, the Spearman rank coefficient was used, as available from the core R package.

Table 6.1 – RC: Correlation between *last year* and chosen features. Universally constant features are marked in italics.

No.	1920–1930		1930–1940		1940–1950		1950–1960		1960–1970	
1.	first half	0.88	first quarter	0.85	first year	0.80	automobile industry	−0.87	other countries	−0.76
2.	floor leader	0.85	second quarter	0.83	international law	−0.76	european countries	−0.82	government officials	−0.75
3.	first quarter	0.84	common stock	0.82	other words	−0.72	democratic leaders	0.81	political parties	−0.70
4.	current year	0.77	current year	0.82	public utility	−0.66	british government	−0.74	european countries	−0.69
5.	farm relief	0.76	*first time*	−0.74	international relations	−0.65	overwhelming majority	0.70	political leaders	−0.66
6.	same period	0.75	tomorrow morning	0.67	national policy	−0.61	*last week*	0.69	other words	−0.64
7.	automobile industry	0.75	first half	0.65	late today	−0.60	british empire	−0.62	open market	0.59
8.	weather conditions	0.73	same period	0.64	near future	−0.57	floor leader	−0.62	low prices	0.57
9.	*same time*	0.72	stock market	0.63	european countries	0.53	american people	−0.60	american government	−0.57
10.	executive session	0.72	low prices	0.63	oil production	0.53	other countries	−0.55	vice president	0.55
11.	whole world	−0.71	american people	−0.62	american people	−0.51	next year	0.54	other nations	−0.53
12.	second quarter	0.69	business conditions	0.61	political parties	−0.51	next month	0.54	present conditions	−0.53
13.	motion picture	0.67	present time	0.60	low prices	0.50	current year	−0.54	third quarter	0.51
14.	vice president	0.66	whole world	−0.58	*last week*	0.49	last summer	0.53	several occasions	−0.51
15.	american people	−0.65	*last week*	0.58	vice president	0.45	disarmament conference	0.53	war debt	−0.50
16.	first year	0.65	law enforcement	0.56	crude oil	0.45	first year	0.52	past week	−0.50
17.	recent years	0.63	good business	0.55	next year	0.45	crude oil	0.51	american people	0.46
18.	american government	−0.62	present indications	0.54	next few	−0.45	present time	−0.46	farm products	0.46
19.	public interest	−0.60	past year	0.53	present conditions	−0.45	recent years	0.45	large majority	−0.45
20.	important factor	0.60	near future	0.52	recent months	0.43	political leaders	−0.45	present time	−0.44
21.	recent weeks	0.57	income tax	0.48	stock market	−0.42	international relations	−0.45	foreign countries	−0.44

The second time window spanning 1930–1940, features various concepts related to businesses and the stock exchange, such *common stock*, *stock market*, *business conditions* and *income tax* as well as a few temporal expressions possibly used in this context, such as *first quarter* and *second quarter*. Interestingly, over the next time span covering 1940–1950, for instance *stock market* goes from being reasonably positively correlated (0.63) to being negatively correlated (-0.42). and other concepts, such as *european countries* and *oil production* take precedence instead. In the next time window (1950–60), the highest rated concepts are negatively correlated with *last year*, this effect becoming even stronger in the very last time frame of 1960–70. Overall, one could interpret this to mean that very different concepts come to be used with *last year* than provided the basis for this set of correlated features. For instance association between *last year* and *stock market* is strong during the years of 1920–1930, but hereafter other concepts start to be used with *last year* instead and previously positive correlation become negative. Certain events, such the surprising Wall Street Crash in 1929 could have

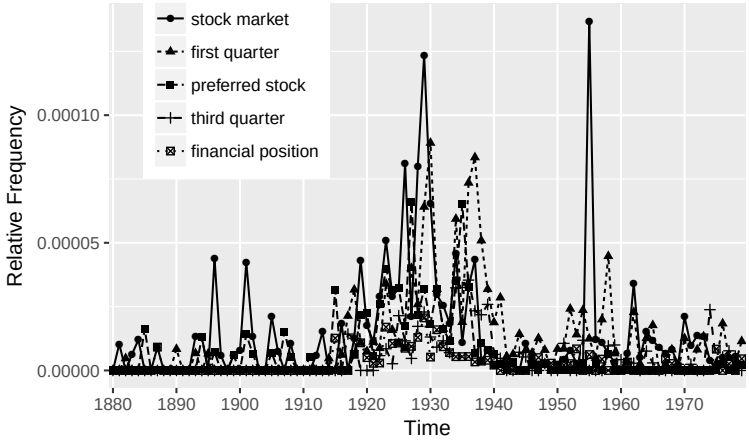


Figure 6.3 – News corpus: relative frequency of items with change-points around 1915–16 and 1945–46.

caused temporal expressions to gain more prominence by creating an atmosphere of immediacy that at least in the news world made the use of temporal expressions more likely. The fact that WWII and the Cold War followed shortly after this event might have kept the temporal dimension palpable.

When examining the list of change-points, including the ones more than 10 years after *last year*, it is noticeable that a few expressions' change-points lie very close together, for instance *stock market*, *preferred stock*, *financial position*, *first quarter* and *third quarter* all change in either the year 1915 or 1916 and then again in 1945 or 1946. Figure 6.3 depicts this overlap in increase after the first change-point and return to initial mean frequency pattern after the second change-point. Another aspect that is noticeable in the results is that various temporal expressions appear in the list of features highly correlated with *last year*. This suggests that temporal expressions in general increased in usage over time with respect to this genre. Fig-

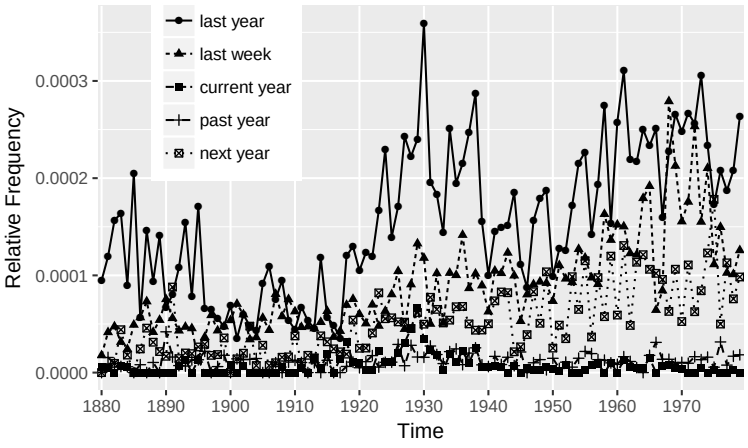


Figure 6.4 – News corpus: relative frequency of temporal expressions.

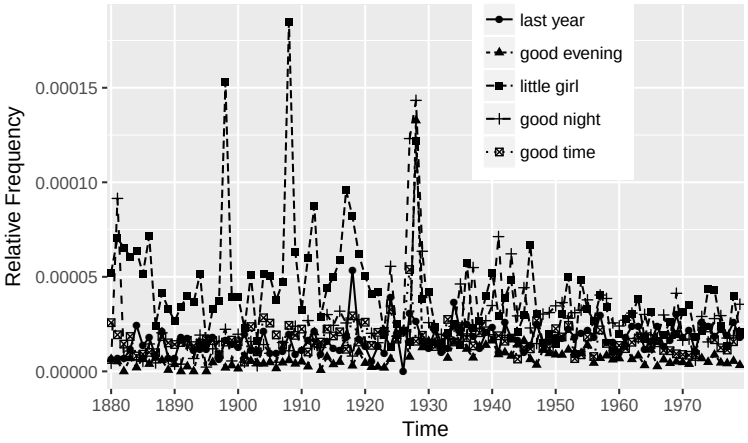


Figure 6.5 – Fiction corpus: relative frequency of highest *last year* correlated features in the fiction genre.

ure 6.4 shows a few of these expressions from Table 6.1. All seem to increase in frequency over time. However, correlation analysis might be a little volatile in that smaller spans of the entire period are not representative of the overall correlation. For this reason, the results are further validated as part of the next section.

6.3.1.2 Validation of Results

The first part of validating the results is to see whether the same effect also exists in the *fiction* genre, which presumably would not be subject to the same influences. Thus, the exact same experiment is repeated, but using the fiction corpus as a basis rather than the news corpus. As a first step, possible change-points on the basis of the new corpus are estimated. Interestingly, the change-point for *last year* in the fiction genre happens earlier, around 1917. Overall, there

Table 6.2 – News corpus: salient words occurring with *last year* in 10 randomly selected sentences for each of the time periods: 1910–1920, 1924–1934, 1950–1960.

News corpus	**salient words occurring with last year**
1910–1920	(primary) election(2), party(2), board of education(2), mexican bullets (1), company (3), director(s)(2), railroad(1), wages(1), shareholders(1), submarine(1), national committee(1)
1924–1934	adjustment bond(1), common stock(2), stock (dividend)(2), sales(1), share(1), corporation(1), net profit(1), minor purchases(1) liquidation(1), dividend rate(1), president(1), preferred dividends(1) (cash) investment(2), congress(1)
1950–1960	tournament(1), basketball coach(1), chicago medical society(1), tax bill(1), (space) administration(2), international agreement(1),wage(1), arbitration(1) net income(1), auto companies(1), production schedules(1), national aeronautics(1), russians pioneer(1)

seems to be a lot less variety in adjective-noun combinations as the partially constant features over 1918–1928 only add up to 27. Of these 27, only 9 are positively correlated with *last year* based on a span of $\pm$15 years around their individual change-points.

However, only *good evening*, *little girl*, *good night*, *good time* and *very well* are actually positively correlated with *last year* over an interval of $\pm$15 years around its own change-point, with the highest correlation being around 0.4. Figure 6.5 shows them side-by-side with *last year*. This suggests that the temporal aspect has not grown as much in importance in this genre and is less closely linked to adjective-noun types as it appears to be the case in the news domain. Validating this in the news data requires actual language samples for co-occurrence of items highly correlated with *last year*. For this, I randomly extracted sentences without replacement containing *last year* and noted what concepts co-occurred in the same sentences. Ten samples were chosen each from before 1923 (1910–1920), immediately after (1924–1934) and again at a later stage (1950–1960).

Table 6.2 shows salient concepts occurring in the same sentence as *last year* for all three time periods. The number in bracket indicates in how many sentences of the ten selected ones the term occurred. The terms occurring in the first time span are mostly related to government elections with some more general political or business topics, such as *election*, *company* and *wages* entering into it as well. The second time span set around the change in *last year* seems to contain almost exclusively stock exchange related news items. The final period, set after the end of WWII contains very mixed samples from sports, to international politics, companies and space programs. Although extracting a few random samples from a large set of texts cannot provide fixed conclusions, these results support the earlier findings of a strong correlation between stock exchange related items and the temporal expression *last year* during a particular time period, where this appears to have dominated the news. In order to see to what extent this effect generalises to other temporal expressions, these need to be analysed separately, which is done in the next section.

Table 6.3 – News corpus: Correlation between *last week* and chosen features. Universally constant features are marked in italics.

No.	1910–1920		1920–1930		1930–1940		1940–1950		1950–1960		1960–1970	
1.	near future	0.83	stock market	0.79	first half	0.64	*first time*	0.69	*last year*	0.69	present conditions	−0.80
2.	long way	0.83	newspaper men	−0.78	near future	0.62	oil industry	0.65	long period	−0.67	city officials	0.60
3.	oil fields	0.82	oil fields	−0.74	common stock	0.59	foreign countries	0.60	financial position	−0.63	former member	−0.50
4.	present conditions	0.80	past year	0.72	*last year*	0.58	american business	0.60	present time	−0.56	financial district	0.47
5.	financial position	0.79	long period	0.71	oil fields	−0.57	crude oil	0.58	american business	0.52	oil industry	0.47
6.	european countries	0.74	past week	0.65	current year	0.54	french government	−0.57	british empire	−0.52	newspaper men	−0.45
7.	preferred stock	0.67	recent months	0.65	first quarter	0.51	other countries	−0.55	*few months*	0.51	crude oil	−0.41
8.	*last year*	0.66	present time	0.62	republican party	0.50	organized labor	0.55	first half	0.50	public utilities	−0.41
9.	other nations	0.66	near future	0.62	financial district	0.50	present time	0.54	low prices	0.47	oil fields	−0.40
10.	several months	0.66	french government	−0.61	other nations	−0.49	other words	−0.52	public utilities	0.47	*few months*	0.37
11.	*few months*	0.64	city officials	0.55	other countries	−0.47	higher prices	0.50	european countries	−0.45	income tax	−0.37
12.	past year	0.60	international relations	0.54	whole world	−0.47	international relations	−0.50	international relations	−0.45	past week	0.36
13.	crude oil	0.60	current year	0.48	stock market	0.45	*last year*	0.49	preferred stock	0.45	recent months	−0.32
14.	large part	−0.57	public utility	0.48	preferred stock	0.45	past week	0.48	foreign countries	−0.45	british empire	0.30
15.	government officials	0.57	republican party	0.47	other members	0.45	oil fields	0.48	world war	−0.45	long period	0.29
16.	next year	0.54	great war	−0.47	other words	−0.44	income tax	0.47	quarterly dividend	0.43	first quarter	0.28
17.	next few	0.54	other countries	0.45	higher prices	0.42	common stock	0.42	lower prices	−0.41	lower prices	0.28
18.	past week	0.51	farm products	−0.45	present conditions	−0.39	whole world	−0.42	home rule	−0.41	european countries	0.27
19.	high school	0.47	recent years	0.44	recent years	−0.38	next year	0.40	current year	−0.39	american people	0.26
20.	long period	0.44	government officials	−0.42	present time	0.35	farm products	0.38	recent years	0.38	*last year*	0.25

Table 6.4 – News corpus: Correlation between *last week* and chosen features based on its second change-point in 1950. Universally constant features are marked in italics.

No.	1940–1950		1950–1960		1970–1980	
1.	last month	0.74	military aid	−0.83	city officials	0.60
2.	technical assistance	−0.65	first year	0.82	*last night*	0.59
3.	foreign ministers	−0.63	news conference	0.72	soviet leaders	0.58
4.	ground forces	0.59	foreign ministers	0.67	political power	−0.58
5.	executive director	0.53	free world	−0.61	news conference	0.57
6.	federal government	−0.52	korean war	−0.60	national committee	0.55
7.	economic aid	−0.52	presidential candidate	0.59	last summer	0.55
8.	*last night*	0.51	power plants	0.59	local government	−0.50
9.	other states	−0.46	atomic weapons	−0.55	national interest	−0.49
10.	news conference	−0.45	next month	0.54	medical care	−0.45
11.	foreign aid	−0.43	communist world	0.54	west german	0.40
12.	last fall	0.41	press secretary	0.52	state government	−0.38
13.	military strength	0.41	*few months*	0.51	foreign ministers	−0.37
14.	military power	−0.38	western nations	0.49	*few months*	0.37
15.	soviet bloc	−0.34	staff members	0.48	military strength	−0.36
16.	state law	−0.33	other areas	0.47	communist world	0.36
17.	state government	−0.32	colored people	0.46	state laws	0.36
18.	staff members	0.32	economic aid	−0.45	staff members	0.35
19.	defense budget	−0.31	defense budget	−0.42	presidential candidate	0.32
20.	korean war	−0.30	last fall	0.41	consumer goods	−0.31

6.3.1.3 Change in Other Temporal Expressions

The previous analysis tentatively suggested that there might be a link between (constant) temporal expressions and partially constant clusters of word sequences relating to political or financial events. In order to investigate this further, two more salient temporal expressions, i.e. *last week* and *next year* are considered using the same analysis as before. Computing change-points for each returns one change-point for *next year* in 1923 and two significant changes for *last week* in 1918 and 1950.[12] First, *last week*'s first change-point in 1918 is examined. Table 6.3 shows the highest pairwise correlations (either negative or positive) between *last week* and each of the 62 partially constant features over smaller intervals of 10 years from 1910 to 1970, where all the universally constant features within are marked in italics. Comparing to the earlier mappings for *last year*, one can observe similar patterns with respect to *last week*'s associations.

Although financial terms appear before 1920, these are most numerous in 1930–1940 and 1940–1950, so longer than for the *last year*. The highest rated terms are similar, e.g. both tables include *common stock*, *stock market*, *preferred stock*, *income tax* and Table 6.3 also contains *financial district* and *net earnings*. This cluster of financial expressions stays positively correlated with *last week* longer than with *last year* – this could mean that there was yet another temporal shift towards more immediacy in reporting, i.e. rather than reporting on shifts on a yearly basis, these might afterwards moved to be expressed on a weekly basis. Interesting topical expressions not directly related to the stock exchange are *oil fields* and *crude oil* – both are positively related to *last week* for the period before and slightly after its first change-point,

[12] This was estimated slightly later than in the previous run for the *last year* comparison.

Table 6.5 – News corpus: salient words occurring with *last week* in 10 randomly selected sentences for each of the time periods: 1908–1918, 1919–1929, 1940–1950, 1970–1980.

News corpus	**salient words occurring with last week**
1908–1918	pacific port, peace agreement, jurymen,county court house, white house,strike, trial,ambassador, london foreign office, state department officials, federal authorities, brokerage business, opening address, army edition, manufacturers' protective association
1919–1929	(federal reserve) bank(2) credit(2) capacity, zoology, lloyd george's offer, prime minister, brigadier general, progressive inflation, broadcast stations, metropolitan museum, long island, dogs' ears, election
1940–1950	citizen, senate amendments, union's convention, communist leadership, government buyer, mayor's office,steel production, furnace repairs, french provisional consultative assembly, storage situation, soviet tactics, british foreign under-secretary, inflationary spiral, automobiles, british mines, average daily output
1970–1980	environmental issue, senatorial candidates,israeli proposals, london correspondent, governor rockefeller, state university, civic center chambers, arab capitals, egypt-israel troop disengagement, arab oil producer, defense lawyers,leftist radicals, union delegates, non-union workers, profit margins, texas wealth and society, domestic auto industry

whereas 1920–1940 shows only negative correlations. However, then association for these and similar expressions are again positive from 1940–1950, and after that mostly negative but for *oil industry*. This could reflect the oil procurement being more topical during certain periods of time, e.g. during WWI and then again during WWII. Table 6.4 shows the associations for *last week*'s second change-point in 1950 based on 71 positively correlated features emerging from the analysis. There is a temporal overlap with Table 6.3 for the period 1940–50. For this period before *last week*'s second change-point, concepts based on its first change-point are more positively related to it. After 1950, the situation is reversed and Table 6.4 holds the higher correlated items. The highest positive associated expressions are very different topically: *news conference*, *foreign ministers*, *power plants*, *communist world* and possibly related to the Cold War that started in 1947. In the next time period, 1970–1980, *soviet leaders* and *west german* are highly correlated with *last week* (*east german* did not quite make the list with 0.31). What is also noticeable is that other temporal expressions are not as numerous as previously, possibly indicating that they were used more independently. Figure 6.4 suggests that *last week* became even more frequent after its second change-point. Table 6.5 shows the result of randomly selecting 10 sentences in which *last week* appears and noting salient noun phrases within. The first time period captures the ten years before its change in 1918: associated concepts include some foreign affairs related expressions, such as *peace agreement*, *ambassador* and *london foreign office* as to be expected with WWI happening during this time.

Table 6.6 – News corpus: Correlation between *next year* and chosen features. Universally constant features are marked in italics.

No.	1920–1930		1930–1940		1940–1950		1950–1960		1960–1970	
1.	*same time*	0.93	international relations	−0.82	international law	−0.77	american people	−0.81	national policy	0.6
2.	public utility	0.89	financial position	0.75	republican party	−0.66	democratic leaders	0.71	other nations	−0.59
3.	lower prices	−0.89	preferred stock	0.66	low prices	0.65	several occasions	0.71	lower prices	0.58
4.	financial position	0.89	recent months	−0.61	financial position	0.61	world peace	−0.70	fiscal year	0.54
5.	purchasing power	−0.86	great importance	−0.58	fiscal year	0.61	recent years	0.64	national defense	0.49
6.	floor leader	0.86	last summer	−0.58	public works	−0.6	overwhelming majority	0.60	long period	0.49
7.	common stock	0.82	fiscal year	0.57	long period	−0.60	public utility	−0.6	other words	−0.47
8.	recent years	0.82	small group	−0.57	tomorrow night	−0.59	interest charges	0.60	net income	0.46
9.	preferred stock	0.82	international law	−0.55	third quarter	−0.56	open market	0.55	*same time*	0.45
10.	current year	0.79	national defense	−0.54	*first time*	0.54	automobile industry	−0.54	present indications	0.40
11.	national defense	0.75	recent weeks	−0.53	american people	−0.51	*last year*	0.54	international relations	−0.40
12.	public officials	0.75	farm products	0.52	political leaders	−0.49	previous year	0.50	executive session	0.38
13.	political leaders	0.75	*same time*	−0.50	public utilities	0.47	law enforcement	0.50	higher prices	0.37
14.	last summer	0.75	*few months*	0.50	several months	0.46	public officials	−0.46	same period	0.36
15.	american legion	−0.75	tomorrow afternoon	0.50	several years	−0.45	international law	−0.45	*last year*	0.35
16.	third quarter	0.74	net income	0.48	great importance	−0.45	office building	0.45	republican party	0.35
17.	second quarter	0.71	interest charges	0.48	*last year*	0.45	whole world	0.45	public officials	0.35
18.	organized labor	−0.71	other members	0.47	weather conditions	−0.43	national defense	−0.45	large majority	−0.35
19.	international relations	−0.71	important factor	−0.46	higher prices	0.43	fiscal year	0.44	stock market	−0.35
20.	tax payments	0.70	floor leader	0.45	important factor	−0.42	other nations	−0.44	public interest	−0.35

Table 6.7 – News corpus: salient words occurring with *next year* in 10 randomly selected sentences for each of the time periods: 1910–1920, 1924–1934, 1950–1960.

News corpus	**salient words occurring with next year**
1910–1920	(republican) party(2), presidential election, railroad companies(2) company, rate, german agricultural council, industrial alcohol, distillers corporation, war department, economy, iron, oil
1924–1934	dividend(2), re-election, candidate, financial position, (common) stock(2) election, business, france, minister of finance, interests, stockholders, budget(2), spending departments, disarmament conference, foreign minister, instalments revenue, income tax, state aid
1950–1960	school head, federal budget, nonmilitary spending programs, consumer and business goods, wages, foreign trade, investment, shareholders, this administration, congress(2), company(2), dividend policy, operating deficit, revenue, chinese communist forces

Other concepts relate to local politics: *state department officials*, *federal authorities* and *county court house*. The time period immediately after the change-point (1919–1929) features some banking concepts (*federal reserve bank*, *credit*) and a variety of seemingly political topics (*lloyd george's offer*, *prime minister*) as well as some more general topics, such as *metropolitan museum* or *zoology*. Both 1940–1950 and 1970–1980's lists seem to be more political, especially with topics concerning the developments overseas and the **threat* of communism. Overall, the results of this random sampling suggest that *last week* might have been more tightly linked to WWI and the ensuing political situations. The last item examined is the change exhibited by *next year*. Like *last year* its change-point is estimated to have occurred in 1923, but unlike the latter it is, although fully constant over the entire reference corpus, actually not universally constant over the news data from 1880–1979. It becomes constant after 1919 and since it was highly correlated with the previous two items in this analysis, it is interesting to consider alongside those expressions. Table 6.6 shows the time period wise highest correlations with *next year* as in the previous cases. The first period (1920–1930) features a number of stock market related terms *common stock*, *preferred stock*, *financial position* and *third quarter* as well as political topics' expressions, e.g. *national defense* or *political leaders*. Both of these level off during the next period (1930–1940) and most previously positively correlated terms become negatively correlated. 1950–1960 probably focused more on European politics, e.g. *democratic leaders*, *open market*, *whole world* with this being somewhat carried forth to the last examined period: 1960–1970. Table 6.7 lists the corresponding salient topics actually exhibited in the same sentences as *next year*. The time period before the change-point (1910–1920) features a lot of political topics: *republican party*, *presidential election* and some broader topics, such as *economy* or *distillers corporation*. Similarly to *last year*'s list, the period from 1924–1925 features a variety of financial expressions (*common stock* or *stockholders*) and various politically related terms: *disarmament conference* and *foreign minister*. The last period (1950–1960) holds but is not limited to more general terms including financial and political expressions, but not being

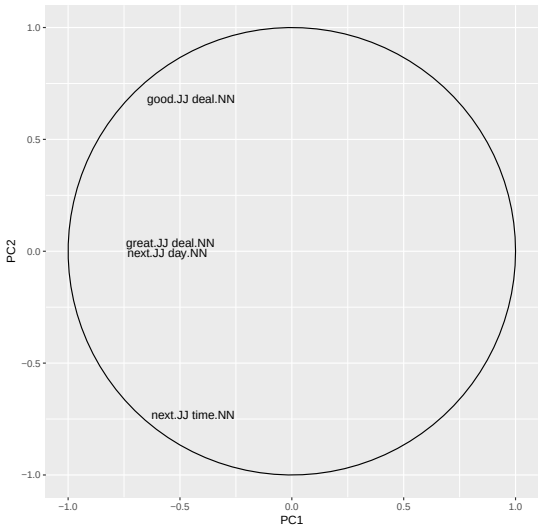


Figure 6.6 – LAC: PCA results for the only four universally constant adjective-noun bigrams.

limited to that. This section has reported an exploratory analysis to investigate the relationship between temporal expressions, such as *last year* and temporally less stable word expressions that appear and disappear over time with respect to the news data. The hypothesis was that these fluctuating words that are more strongly connected to current events would somewhat influence the rise in frequency of more stable concepts, such as temporal expressions. The results suggest that there might indeed be a connection between temporal expressions and clusters of words linked to historical events, such as the Wall Street Crash or WWI/WWII. However, while stock market related words are only constant and very frequent for a limited time frame, *last year* and other temporal expressions remain frequent. This could be due to temporal aspect in news language having become more important after 1923, having gathered momentum through events, such as the stock market crash and then remained to stay. The parallel analysis of temporal expressions in fiction data at the same time seems to confirm this insofar as this effect is not found with the same strength in fiction data. Based on the language sample analysis that appears to support the change-point and correlation analysis, all three temporal expressions are continued to be used in various different contexts and possibly more varied than before 1923.

The results presented need validation from historians, especially with respect to events in 1923 that could have caused temporal expressions to become more frequent. The type of analysis done here shows changes in words' relative frequency patterns that could reflect political or cultural changes. In this, the analysis is at the mercy of the sampling of the newspaper corpus that although balanced over different sources is not impervious to other external factors that could influence the language samples. For instance, by the mid-1920s, the businessman William Randolph Hearst had acquired 28 newspapers[13], that consequently have been subject

[13]The list included: *The San Francisco Examiner*, *The New York Journal*, *Los Angeles Examiner*, *Boston Ameri-*

to same editorial decisions, distorting linguists' perception of what *news* language was representative for that time.[14]

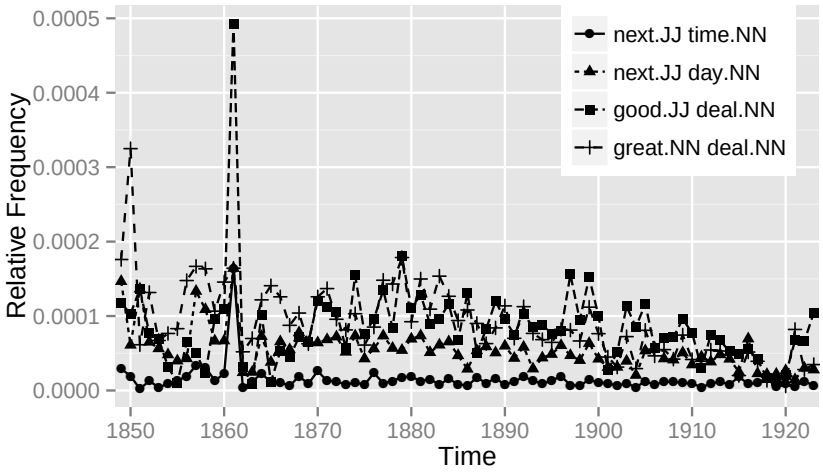


Figure 6.7 – LAC: relative frequency of four universally constant features.

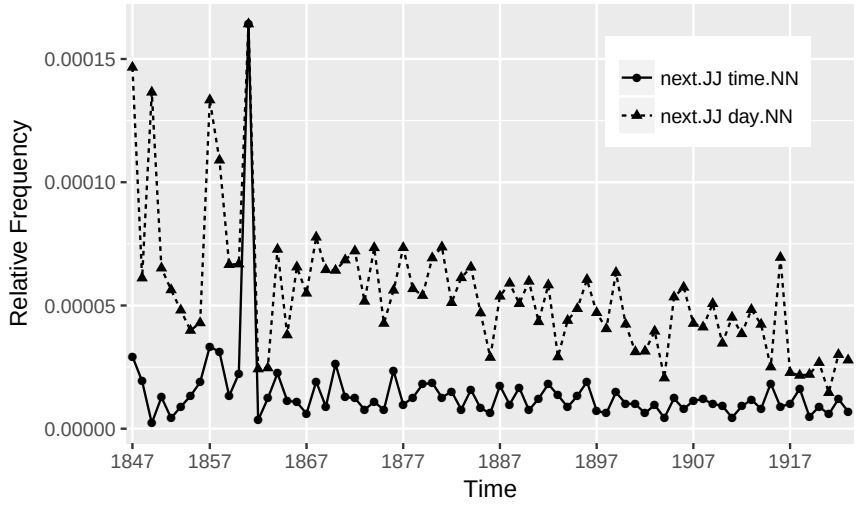


Figure 6.8 – LAC: relative frequency of two universally constant temporal features.

6.3.2 Temporal Expressions in Literary Style

The analysis in Section 6.3.1 has shown that change in adjective-noun and noun-noun bigram types could differ considerably based on genre. More specifically, the exploration of

can, *Chicago Examiner*, *Detroit Times*, *Washington Times* and *Washington Herald*.

[14]This newspaper collection does not appear to contain a Hearst-owned publication, but might still have been influenced by those rivalling newspapers.

Table 6.8 – LAC: Correlation between *next day* and chosen features. Universally constant features are marked in italics.

No.	1875–1885		1885–1895		1895–1905		1905–1915	
1.	pale face	−0.62	many ways	−0.79	*great deal*	0.87	good looks	0.72
2.	many people	0.60	poor man	0.77	other words	−0.86	next year	0.67
3.	other girls	0.57	last year	0.74	next morning	0.83	little while	0.65
4.	next year	0.56	little while	−0.68	*next time*	0.82	own heart	−0.65
5.	hard work	−0.55	good care	0.62	last year	0.68	blue eyes	−0.62
6.	few weeks	−0.52	few weeks	−0.61	many ways	−0.66	next room	0.59
7.	last year	−0.48	next year	0.60	poor man	0.65	whole matter	0.58
8.	next room	−0.48	old times	−0.55	little while	0.64	other woman	−0.55
9.	last moment	−0.46	young girls	−0.50	good looks	0.64	*great deal*	0.55
10.	*great deal*	0.44	worth while	0.50	good sense	0.59	poor man	0.54
11.	strange thing	0.43	such cases	−0.49	little money	0.53	hard work	0.54
12.	poor man	−0.40	own mind	−0.47	next year	0.48	other girls	0.53
13.	such cases	−0.39	strange thing	0.47	quiet way	0.45	little money	−0.51
14.	short laugh	0.37	few people	0.45	few things	−0.44	good taste	0.5
15.	something new	−0.35	young fellow	−0.43	young people	0.39	own mind	−0.49
16.	most people	0.33	good taste	0.40	few weeks	0.39	most people	0.46
17.	good man	−0.33	old age	−0.35	many people	0.38	faint smile	−0.45
18.	many men	0.31	own life	−0.35	blue eyes	−0.38	few things	−0.45
19.	blue eyes	−0.29	first moment	0.35	other man	0.38	old times	0.41
20.	few people	0.26	other words	0.34	old times	0.37	whole affair	−0.38

co-occurrence relationships between universally constant and partially constant features with focus on news data in Section 6.3.1.2 suggests that temporal expressions did not change in the same fashion in the fiction corpus. Nevertheless in this section, salient temporal expressions are analysed with respect to the literary authors' corpus in order to gain some intuition about the way these features change for literary authors and thereby investigate additional possible differences to the news corpus.[15]

As a first step, again, PCA is applied to the joint literary authors' universally constant features to discover features with common variance patterns over time. Extracting universally constant features over the corpus only yields four adjective-noun combinations: *next day*, *great deal*, *good deal* and *next time*.[16] Subjecting these four expressions to a PCA analysis as was done previously yields the mapping shown in Figure 6.6. Figure 6.7 shows the relative frequency for all four features over time and Figure 6.8 shows a close-up of only the two temporal expressions: *next time* and *next day*, which both seem to decline over time.

Conducting a change-point analysis estimates that the expression *next day* has a change-point in 1885 and from Figure 6.8, one can observe that frequencies seem to be declining after that.[17] The corresponding partially constant adjective-noun/noun-noun features that correlate

[15]This analysis is based on the joint authors' corpus without processing each author individually. This means that feature counts were solely based on the year some text originated in without reference to the authorial source. The change-points based on their averaged data are either the same or vary 5–10 years, thus there may be a slight shift in feature associations when computed that way. The results in this section suggest that the change in temporal features is less likely to have originated in changing patterns of non-constant features indicating that causes might be more subtle and possibly related to a shift in narrative style rendering exact change-points less important.

[16]It is noteworthy that both *great deal* and *good deal* emerged as highly salient features for the news data as well (see beginning of section 6.3.1).

[17]Using a different type of averaging over the corpus, ie. smoothed over two years around the year in question, as used in Section 5.2.3 results in an estimate 7 years later (1892).

Table 6.9 – LAC: salient words occurring with *next day* in 10 randomly selected sentences for each of the time periods: 1875–1885, 1885–1895, 1913–1923.

News corpus	**salient words occurring with next day**
1875–1885	flannel bag, cold place, dead tree, haunted house, good weather, cold water, coffee-house, cigar,same hour, pretty salon, bridal raiment, nigger, other people
1885–1895	local worthies, vague identity,affair, society, neat speech, delicate regret, incredible air, modern personage, strange new house, rodeo, coast, headquarters, engagements
1913–1923	gentle old manager, delightful friend last exercises, little church, funerals, sermon, atonement, clothes-line, excitement, solitary place, concert

Table 6.10 – LAC: Correlation between *next time* and chosen features. Universally constant features are marked in italics.

No.	1871–1881		1881–1891		1891–1901		1901–1911	
1.	own heart	−0.72	old times	−0.65	old house	−0.68	many people	0.82
2.	good man	−0.55	next room	0.62	white face	−0.59	young people	0.8
3.	white face	−0.54	hard work	0.57	old times	−0.56	last year	0.77
4.	little while	−0.53	next year	0.49	good while	0.55	whole affair	0.75
5.	faint smile	−0.53	last moment	0.47	next day	0.55	other words	−0.70
6.	good care	−0.47	many men	0.43	most people	0.54	good sense	0.69
7.	other girls	0.46	worth while	0.42	many ways	−0.52	forty years	0.66
8.	best thing	0.37	other woman	0.42	such matters	0.5	best thing	0.65
9.	whole affair	−0.36	many people	−0.39	good man	0.49	old maid	0.65
10.	open door	−0.34	old age	−0.39	other words	−0.48	good opinion	0.64
11.	most people	−0.33	own heart	0.35	faint smile	0.48	other girls	0.56
12.	many people	0.32	young fellow	0.32	young people	−0.46	few weeks	0.54
13.	few weeks	−0.32	other girls	0.32	open door	0.41	strange thing	−0.54
14.	own name	0.32	own life	−0.31	own heart	0.41	next year	0.53
15.	young people	0.28	whole affair	−0.31	own sake	0.36	next morning	0.50
16.	little money	−0.28	young girls	0.30	other woman	−0.34	old house	0.45
17.	strange thing	0.27	faint smile	0.30	due time	0.33	own name	0.43
18.	few people	0.25	strange thing	0.30	young fellow	−0.32	hard work	0.42
19.	same thing	0.25	old maid	0.27	young girl	0.31	own sake	0.42
20.	next year	−0.25	next morning	0.25	whole affair	0.31	next day	0.39

strongest with *next day* for several particular time spans are shown in Table 6.8. The concepts for the first period (1875–1885), 10 years before the change-point, are shown in the first column. The highest positively associated concepts are *many people*, *other girls* and *next year*. For the period of 10 years after the change-point covering the years 1885–1895, previously negatively correlated concepts are highly positively correlated: *poor man* and *last year*. From 1895–1905, *great deal* is the highest associated feature and new features include *good looks* and *good sense*. The phrase *blue eyes* becomes more highly negatively correlated with time. Table 6.9 shows salient expressions occurring in the same sentences as *next day*, which support the correlation analysis insofar that there is no visible shift in topic associated with the way the expression is used suggesting that causes for its slow decline might be more subtle. Figure 6.9 shows *next day* for the reference fiction corpus and the literary authors' corpus side-by-side, indicating that the authors' corpus seems to follow the same general trend for that feature. In

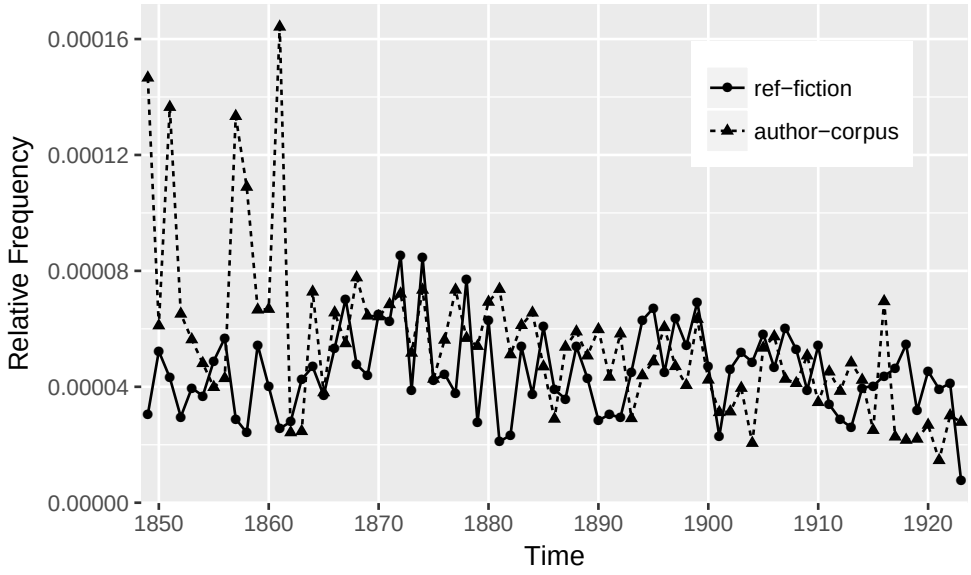


Figure 6.9 – LAC and RC fiction for the feature *next day*.

order to see whether this pattern extends to the other temporal expression, the same change-point analysis is carried out for *next time*. The change-point is estimated to have happened in 1881 and Table 6.10 shows the highest correlated features for *next time*. Similarly to *next day*'s correlations, intuitive topics do not immediately come to mind when examining the list and association changes over time. For instance, *young fellow* and *young people* appear frequently in all four periods, but positive and negative associations do not seem to be related to the change-point. However, this could represent a change in narrative structure that is not connected to the introduction or loss of other expressions.

As a final exercise, to compare to the most salient temporal expression for the *news* data, *last year* is examined for the LAC. The expression *last year* is not constant over this corpus, indicating a possibly different distribution in comparison to the RC news data. The estimated change-point is the year 1869, which lies earlier than the one estimated for the RC fiction genre (1917) – however there could be more than one change-point associated with a feature, and as has been observed with *last week*, this need not indicate a change in frequency direction, i.e. a feature can further increase to a significantly different level to the one before. Figure 6.10 shows the relative frequency of *last year* over the LAC and the RC fiction data; for the reference fiction corpus it rises to a new level after 1920 and appears to remain there. As this literary authors corpus does not have any data points after 1923, it is not entirely certain whether their data would follow the same trend.

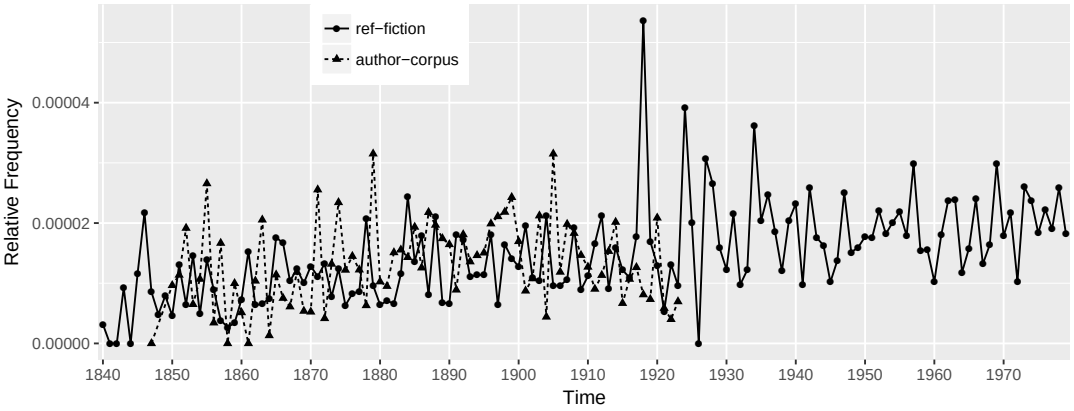


Figure 6.10 – LAC and RC: relative frequency of *last year*.

Table 6.11 – LAC: Correlation between *last year* and chosen features. Universally constant features are marked in italics.

No.	1869–1879		1879–1889		1889–1899		1899–1909	
1.	old house	0.70	last moment	0.82	old maid	0.79	*next time*	0.80
2.	pretty girl	0.69	young girl	0.65	next morning	0.77	other girls	0.80
3.	young girls	0.67	many men	0.64	tall figure	−0.75	own heart	−0.79
4.	old maid	0.61	little child	−0.53	other girls	0.70	open door	−0.77
5.	such matters	0.57	old house	−0.51	whole affair	−0.68	little money	0.74
6.	young girl	0.55	white face	0.47	own sake	−0.64	next morning	0.70
7.	own life	0.51	other words	0.46	own mother	0.58	next year	0.66
8.	other woman	0.49	hard work	0.41	young people	0.54	good opinion	0.66
9.	little while	−0.48	tall figure	0.39	due time	−0.50	*next day*	0.64
10.	good care	0.47	*next day*	−0.38	*next day*	0.49	old maid	0.64
11.	best thing	0.46	same thing	0.37	same thing	−0.45	other words	−0.61
12.	last moment	0.46	little while	−0.35	last moment	0.42	many ways	−0.56
13.	few people	0.45	good while	−0.35	good man	−0.40	own accord	−0.55
14.	own mind	0.44	due time	0.34	own heart	−0.38	white face	−0.55
15.	poor mother	−0.42	next morning	−0.32	many men	−0.37	own mother	0.55
16.	white face	−0.40	other woman	0.32	good care	0.35	small table	0.53
17.	old times	0.39	old age	0.31	young girls	−0.35	cold water	0.49
18.	little money	0.36	young people	0.29	old age	0.33	few weeks	0.49
19.	good opinion	−0.35	best thing	−0.29	next year	0.31	young people	0.48
20.	hard work	0.35	such matters	0.28	little child	0.31	few people	0.46

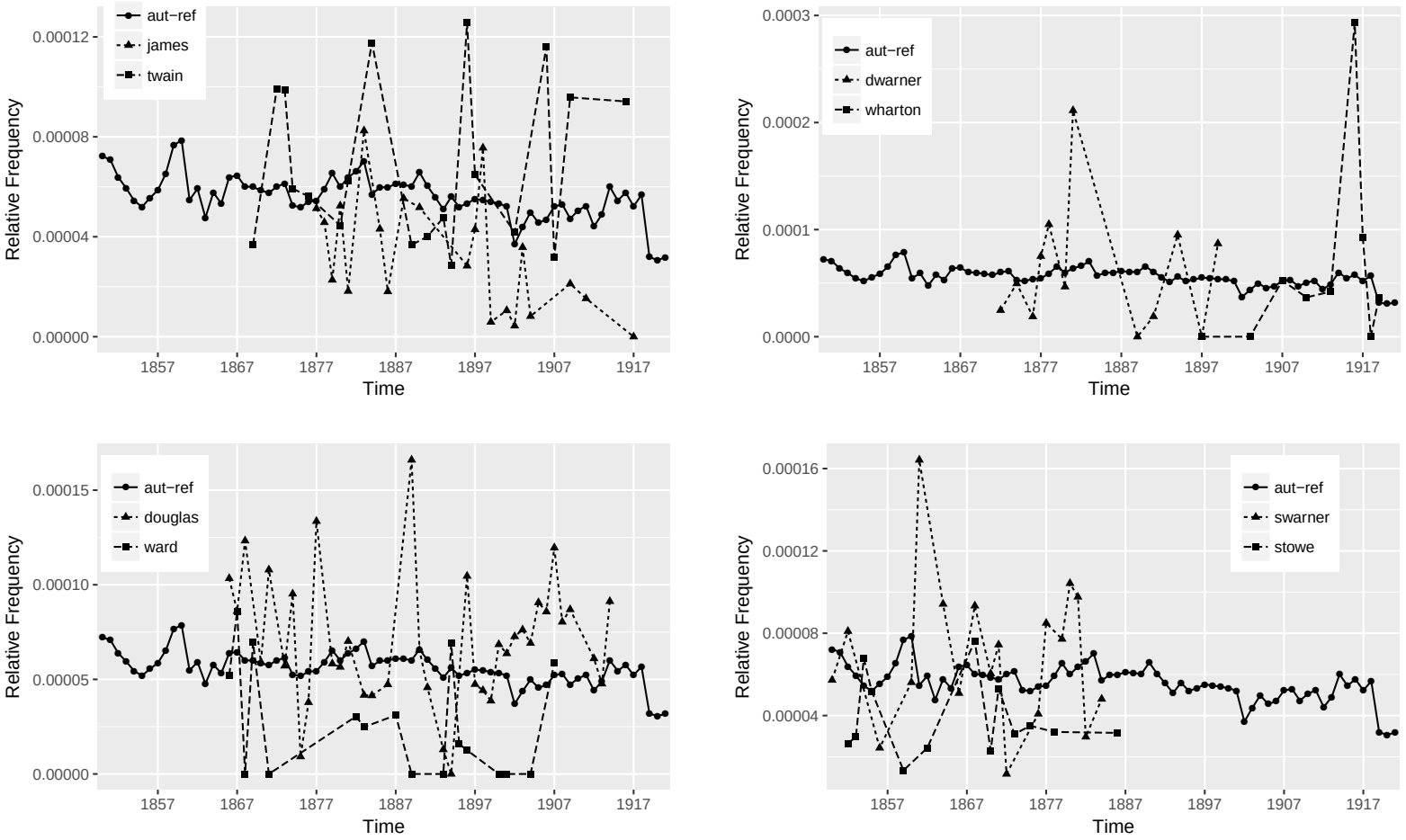


Figure 6.11 – LAC: relative frequency of *next day* for: Twain, James, Charles Dudley Warner, Wharton, Douglas, Ward, Susan Warner and Stowe alongside the average over all authors (ARC).

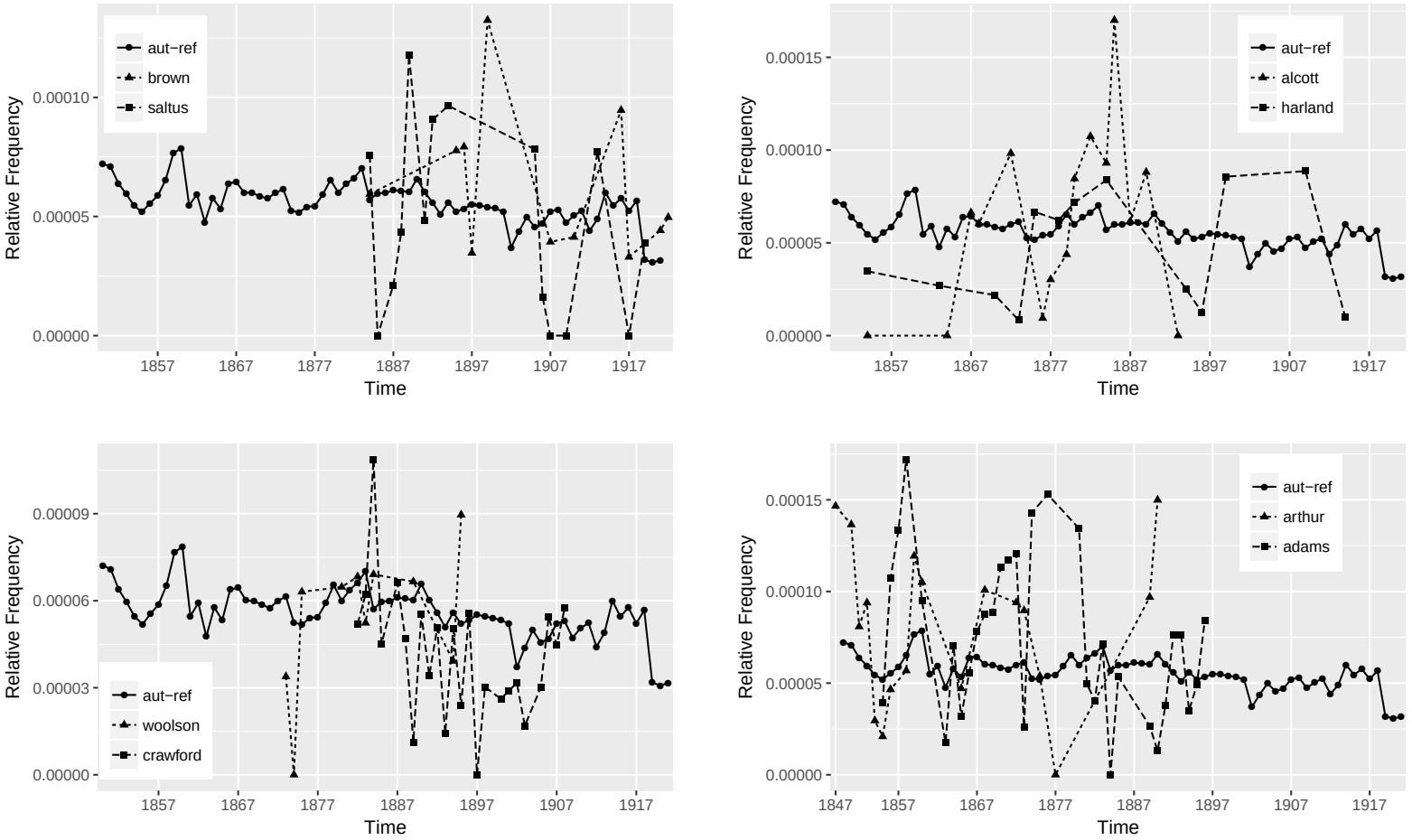


Figure 6.12 – LAC: relative frequency of *next day* for Brown, Saltus, Alcott, Harland, Woolson, Crawford, Arthur and Adams alongside the average over all authors (ARC).

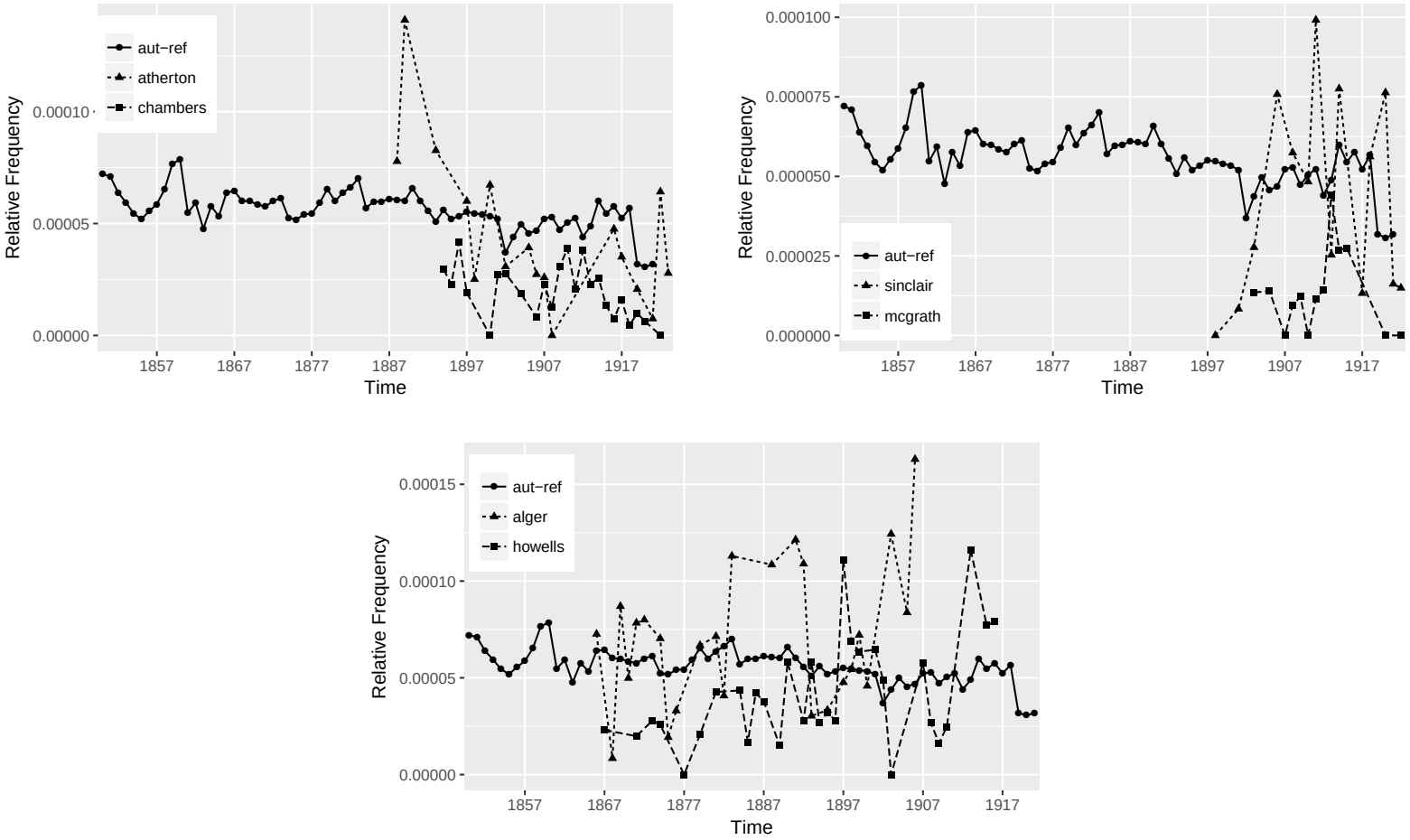


Figure 6.13 – LAC: relative frequency of *next day* for Atherton, Chambers, Sinclair, McGrath, Alger and Howells alongside the average over all authors (ARC).

Table 6.11 shows the highest correlations with *last year* for the literary authors' corpus. Similarly to the previous analyses, there does not seem to be a strong change in topic after the change-point. Therefore, the results are somewhat inconclusive in that although there does exist measurable change in the literary authors as well as the RC fiction section, it does not seem to relate to other expressions that change around the same time in a very obvious fashion. This could have several different reasons, one of which might be that individual literary authors as opposed to newspaper journalists each have a very strong unique signal and if analysed in unison these may become distorted and do not offer grounds for solid interpretation. This could only be ascertained by considering each author individually for different types of word classes. The results, however, do suggest that news language does change somewhat differently from literary language with respect to temporal expressions. What is noticeable is that change-points for *last year* based on the literary corpora were estimated to have happened earlier than in the news domain, suggesting that not only might literary authors be more in control of their style and less susceptible to outside changes, but also that they in fact may have influenced change in usage, even if it then took on a development different from their own. One side question interesting to explore is whether there were significant differences in style between authors for these features and specifically whether Mark Twain and Henry James displayed different stylistic behaviour from their peers for these features.

Individual Author Change Figure 6.11, Figure 6.12 and Figure 6.13 show the features *next day* for each author in the corpus separately, grouping authors together in pairs. For each author, the smoothed average over all authors is added for reference.[18] In general, individual authors' development for the feature *next day* vary along three different dimensions: the level of relative frequency with respect to the overall average, i.e. above, below, or varying around the average; whether there is a discernible trend in the feature over time; and the level of variance over time. Authors can be somewhat grouped according to these characteristics. For instance, authors with a similar downward pattern as the ARC as well as the RC fiction part are: Henry James, Susan Warner and Harriet Beecher Stowe (Figure 6.11), Alice Brown, William Taylor Adams and Francis Crawford (Figure 6.12), Gertrude Atherton and Robert Chambers (Figure 6.13), where for instance James, Stowe and Crawford have a relative frequency level below, Brown above and Warner around the overall author average. The two authors Horatio Alger and William Dean Howells actually display somewhat of an increase over time (Figure 6.13). However, most of the remaining authors show no clear pattern, i.e. Mark Twain, Charles Dudley Warner, Edith Wharton, Amanda Minnie Douglas, Elizabeth Ward in Figure 6.11, and Edgar Saltus, Louisa May Alcott, Marion Harland, Constance Fenimore Woolson and Timothy Shay Arthur in Figure 6.12, and Upton Sinclair and Harold McGrath in Figure 6.13. There are differences with respect to frequency level compared to variation and the average, where for instance Twain, Charles Dudley Warner and Woolson show quite a lot of amplitude changes.

[18] This average is computed as outlined in Section 5.2.3.

Given that Mark Twain does not display development for the feature over time and Henry James shows a similar trend of decline for this feature as the ARC, Horatio Alger and William Dean Howells are the only ones that emerge as having a somewhat different style change at least for the feature examined. This does not imply that Twain and James do not possess similarly opposing features, but rather that their style change may not be remarkably different from their fellow authors in spite of what their having been somewhat in the limelight might have suggested otherwise.

6.4 Discussion

The previous section examined a different type of change, namely change that occurs more suddenly and that could possibly be linked to frequency changes in other clusters of words. For this, two different types of language genre were considered: *news* and *fiction*. The type of change focused on for this work was changes in topics likely to be caused by important events. The method proposed here returned good results in that it was able to identify relations between universally constant and partially constant expressions with respect to the news corpus that tentatively suggest frequency relationships between universally constant features and emerging and disappearing clusters of expressions that could help understand how features that are *always* there change in their frequency patterns through less constant features. The present analysis has yet to be tested and extended for the use of other word types and the relationship between these, for instance by considering how nouns and verbs interact over time. Applying the same analysis to fiction data has not yielded similarly intuitive results and more detailed analysis and external validation through literary experts would be needed to fully interpret the present results. However, the results based on the fiction corpus support those of the news corpus as one would expect word cluster changes to be more pronounced in that domain due to the reporting nature of writing, where the choice of topic would be rather outside the journalists' control. Additionally, these findings may indicate that literary authors could have been less susceptible to popular development pervading other genres, having more conscious control over their own style. To what extent this would not be true for non-professional writers is difficult to determine.

This analysis using change-points adds to a simpler relative frequency detection approach by considering the uncertainties associated with the predictions. Although without having conducted a semantic change analysis, one cannot be entirely certain that this change has not been caused by a shift in semantics, however, the possible semantic space of temporal expressions could be seen as being more limited than that for regular common nouns or adjectives. In fact, these temporal expressions might semantically be closer to function types than to content types, in spite of belonging to the latter word class. In a sense, temporal adverbs are similar to prepositions, only anchored in time rather than in space and consequently there might be less room for reinterpretation of their meaning. This analysis, therefore, also somewhat challenges the open-

class/closed-class view of features, suggesting that this might be insufficient when observed through the lens of temporal text representation. While something might bear the label of content word, it could in fact be closer to a function word, behaving and being affected by similar factors, such as regular occurrence in different contexts. An example of this would be phrasal verbs that through their many applications are very likely to appear more frequently. Features are usually classified along one or more different dimensions, such as membership of either an open-class or closed-class or according to the frequency strata, i.e. frequent, medium-frequent and rare, they belong to. However, even though these classifications are often represented as categories suggesting there is a clear boundary between, for instance what is frequent and infrequent, a continuous representation, especially considering temporal effects would occasionally seem more suited, rendering another more comprehensive, less binary definition necessary, in short adopting "...the view that each individual word is positioned on a continuum ranging from fully grammatical in nature to fully content-bearing" [Halteren and Oostdijk, 2015, p.207].

Overall, the contribution of this work was the introduction of methods for analysis of relationships between words solely based on co-occurrence patterns, rendering the task of establishing relationships between words more difficult than, for instance by analysis based on multi-word expressions.

6.5 Conclusion

In essence, this work has been exploratory trying to connect groups of words that might not occur close to each other within texts making their relatedness less tangible. Although additional work is needed to further support the findings, the results indicate that words or expressions that are stable in occurrence, might be rather volatile with respect to their relative frequency distribution. As temporal expressions have fewer semantic associations, they might depend more strongly on features that do.

The results obtained are tentative and in order to claim an increase of temporal expressions possibly related to certain historical events, one needs to show this effect to hold for a variety of temporal expressions as well as exclude any possible semantic shift. This also needs validation from historians to interpret and relate the results to historical and cultural changes in or around the measured change-points. Particular language usage and change therein can reflect shifts in society and general opinion, adding a more subtle basis for interpretation of past events.

Chapter 7

Conclusion

This thesis has examined language change in literary authors taking into account general underlying language change through examining an aligned reference corpus. As part of this, methods for the analysis of stylistic change in authors have been introduced to specifically consider aspects of ageing, general stylistic change and analysis and interpretations of sudden changes in style.

The question of salient aspects of linguistic change has been considered in Chapter 4. The features examined here did not provide evidence of linguistic ageing in the literary authors, however there was some evidence for background language interference.

Further, as concerns salient aspects of general stylistic change, Chapter 5 has provided methods for detection as well as models that estimate the influence of underlying language change. When literary authors are analysed as a group, more general features, such as character trigram and tetragrams, as well as syntactic ngrams are likely to be the most telling features of literary change. Syntactic word ngrams and stem ngrams seem to vary more strongly depending on author and may be more useful for individual analyses.

Chapter 6 considered the question of more abrupt changes in frequency of constant features and how this related to other non-constant features. Temporal expressions in news data have emerged as interesting candidates of sudden changes likely caused by political events. The effects identified with respect to the news data do not exist with the same strength in the literary authors.

Finally, the question of whether James' and Twain' change is markedly different with respect to other authors composing works at the same time has been addressed in Chapter 4, Chapter 5, and Chapter 6. In spite of them having received considerable attention through other analyses, the analyses here could not detect anything that set them distinctly apart from their contemporaries.

This thesis has presented methods for quantification of literary style change in relation to background language, in the process identifying elements of style change.

Appendix A

Literary Authors: Data sets

The following tables show the data collection information for each author. Of highest priority was to obtain a version of a text with a time stamp closest to its first publication date, however, the quality of the text, i.e. level of noise introduced by the OCR process was also an essential factor in this. In some cases, although first edition texts were available, the number of errors seemed to outweigh the benefit of including them and a later version was chosen instead.

Each table's first column shows the first publication date of a work ('First PUB Date'), i.e. the first time it had been published according to different online sources. Sometimes, information as to exact publication dates was ambiguous and those cases are marked with a '?'. The second column, 'Assigned Date', refers to the date assigned to the text, this could be the first publication date or otherwise a date earlier or later. This column directly corresponds to the 'Dating Source' column that indicates which source the assigned date is based on. There are different types: 'PREF' for preface, 'PUB' for publication date, 'CP' for copyright and 'EAC' stands for a note sometimes appearing in the preamble of the book, i.e. "Entered according to Act of Congress, in the year ...". These four different types are not mutually exclusive and sometimes did appear alongside each other. Although, these dates would occasionally be perfectly aligned, very often copyright and publication dates differed by a year or two, and sometimes even by 5–10 years. The objective for dating these texts was to date them as accurately as possible with respect to the year they were composed in and for this reason, whenever possible the first formal date was used, even if it was only a dated preface. The rationale behind this is that a text is likely to always be more similar to the language of the author at composition time, even if it undergoes some changes for republication afterwards. The only exception made to this was when a text underwent major revisions, as would be indicated in for instance a preface. These texts are then assigned the new publication date and marked with a '*R*' to indicate substantial revisions. The fourth column 'CP Dates' and fifth column 'PUB Dates' list explicit copyright and publication dates. These were not always allocated to the same person, although this is not distinguished here. The last column 'Publisher' lists information as to the publishing company, whenever available. Further superscripts mark properties of the specific work,

i.e. '*' means that a work has remained unfinished, '†' indicates that the work was published posthumously, '*R*' stands for revisions and '*C*' indicates that the work was written in collaboration with another author, where the current author was listed as 'first author'. Collaborations were the exception here and would ideally not be included at all, but were added sometimes to augment a sparser timeline, thus judging it to be more beneficial to include them, even if this inclusion might slightly distort later results.

Table A.1 – Collected works for Louisa May Alcott.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
Flower Fables	1854	1854	PREF			
Moods	1864	1864	EAC	1864		Loring, Publisher
On Picket Duty	1864	1864	PUB		1864	
The Abbot's Ghost	1867	1867	PUB		1867	
Little Women	1868	1868	EAC	EAC: 1868,1869, CP: 1880,1896		Little, Brown, and Company
Shawl Straps	1872	1872	PREF		1895	Sampson Low, Marston & Company
Rose in Bloom	1876	1876	PREF			
A Modern Mephistopheles	1877	1877	CP	1877, 1889	1889	Roberts Brothers, Publishers
Aunt Joe's Scrapbag Vol.5	1879	1879	CP	1879	1880	John Wilson and Son, Cambridge
Jack and Jill	1880	1880	PREF			
Aunt Joe's Scrapbag Vol.6	1882?	1882	CP/PUB	1882	1882	Roberts Brothers
Spinning-Wheel Stories	1884	1884	CP	1884	1902	Little, Brown, and Company
Lulu's Library	1885?	1885	CP	1885	1886	Boston: Roberts Brothers
A Garland For Girls	1887	1887	PREF			
Lulu's Library Vol.3†	1889	1889	CP/PUB	1889	1889	Roberts Brothers
Comic Tragedies†	1893	1893	CP/PUB	1893	1893	Roberts Brothers

'†' indicates posthumously published works.

Table A.2 – Collected works for Gertrude Atherton.

Title	First PUB Date	Assigned Dating Date	Source	CP Dates	PUB Dates	Publisher
What Dreams May Come	1888	1888	CP	1888		Belford, Clarke and Company
Hermia Suydam	1889	1889	CP			The Current Literature Publishing Co.
The Doomswoman	1892	1893	CP	1893		J. Selwin Tait & Sons
Patience Sparhawk and Her Times	1897	1895	CP	1895,1897	1897	John Lane: The Bodley Head
The Valient Runways	1898	1898	PUB		1898	Dodd, Mead and Company
The Californians	1898	1898	PUB		1898	John Lane: The Bodley Head
The Senator North	1900	1900	CP		1900	John Lane: The Bodley Head
The Conqueror	1902	1902	PUB		1902,1904	The Macmillan Company
The Bell in the Fog and Other Stories	1905	1905	PUB		1905	Harper & Brothers
The Travelling Thirds	1905	1905	CP/PUB	1905	1905	Harper & Brothers, Publishers
Rezanov	1906	1906	PUB/CP	1906	1906	The Authors and Newspapers Association
Ancestors	1907	1907	CP/PUB	1907	1907	Harper & Brothers
The Gorgeous Isle	1908	1908	CP/PUB	1908	1908	Doubleday, Page & Company
Mrs Balfame	1916	1916	CP	1916		Frederick A. Stokes Company
The Living Present	1917	1917	CP	1917		Frederick A. Stokes Company
The Avalanche	1919	1919	PUB		1919	
Sisters-In-Law	1921	1921	CP	1921		Frederick A. Stokes Company
Sleeping Fires	1922	1922	CP	1922		Frederick A. Stokes Company
Black Oxen	1923	1923	CP	1923		A. L. Burt Company

Table A.3 – Collected works for Alice Brown.

Title	First PUB Date	Assigned Dating Date	Assigned Dating Source	CP Dates	PUB Dates	Publisher
Stratford by the Sea	1884	1884	CP/PUB	1884	1884	Henry Holt and Company
Meadow-Grass: Tales of New England Life	1895	1895	PUB			
By Oak and Thorn	1896	1896	CP/PUB	1884	1884	Houghton, Mifflin and Company
The Day of his Youth	1897	1897	CP	1897		Houghton, Mifflin and Company
Tiverton Tales	1899	1899	CP/PUB	1899	1899	Houghton, Mifflin and Company
Rose MacLeod	1907	1907	CP	1907,1908	1908	Grosset & Dunlap
John Winterbourne's Family	1910	1910	CP/PUB	1910	1910	Houghton Mifflin Company
Country Neigbours	1910	1910	CP/PUB	1910	1910	Houghton Mifflin Company
The Prisoner	1916	1916	PUB/CP	1916	1916	The Macmillan Company
Bromley Neighbourhood	1917	1917	CP	1917		
Louise Imogen Guiney	1921	1921	CP/PUB	1921	1921	The Macmillan Company
Old Crow	1922	1922	CP/PUB	1922	1922	The Macmillan Company

Table A.4 – Collected works for Amanda Minnie Douglas.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
In Trust	1866	1866	EAC	1866	1872	Lee and Shepard, Publishers
Stephen Dane	1867	1867	EAC/PUB	1867	1867	Lee and Shepard, Publishers
Syndie Adriance	1868	1868	EAC	1868	1869	Lee and Shepard, Publishers
Kathie's Three Wishes	1870	1870	EAC	1870(EAC),1898(CP)		Lee and Shepard, Publishers
Kathie's Aunt Ruth	1870	1870	EAC	1870	1883	Lee and Shepard, Publishers
With Fate against Him	1870	1870	EAC/PUB	1870	1870	Sheldon & Company
Kathie's Soldiers	1870	1971	EAC	EAC:1871 CP:1899		Lothrop, Lee & Shepard Co.
Harvest Days	1870	1871	EAC	EAC:1871 CP:1899		Lothrop, Lee & Shepard Co.
Home Nook	1873	1873	EAC	EAC:1973 CP:1901		Lothrop, Lee & Shepard Co.
The Old Woman Who Lived in a Shoe	1874	1874	EAC	1874		Lee and Shepard
Seven Daughters	1874	1874	EAC	1874		Lee and Shepard, Publishers
Drift Asunder	1875?	1875	EAC	1875		William F. Gill & Co.
Nelly Kinnard's Kingdom	1876	1876	EAC/PUB	1876	1876	Lee and Shepard, Publishers
From Hand to Mouth	1877	1877	CP	1877,1905		Lee and Shepard, Publishers
Hope Mills	1879	1879	CP	1879		Lee and Shepard, Publishers
Lost in a Great City	1880	1880	CP	1880,1908		Lothrop, Lee & Shepard Co.
A Woman's Inheritance	1881?	1881	PREF	1886		Lee and Shepard, Publishers
Floyd Grandon's Honor	1883	1883	CP	1883	1899	Lee and Shepard, Publishers
Whom Kathie Married	1883	1883	CP/PUB	1883	1883	Lee and Shepard, Publishers
Out of the Wreck	1884	1884	CP	1884		Lee & Shepard Co.
Foes of Her Household	1886	1886	CP	1886		Lothrop, Lee & Shepard Co.
A Modern Adam and Eve	1888	1889	PUB		1889	Lee and Shepard Publishers
A Little Girl in Old Philadelphia	1899?	1890	CP	1890		A. L. Burt Company
The Heirs of Bradley House	1891	1891	PREF			Lee and Shepard Publishers
In the King's Country	1893	1893	CP	1893, 1894	1894	Lee and Shepard Publishers
In Wild Rose Time	1894?	1894	CP/PREF	1894		Lothrop, Lee & Shepard Co.
A Little Girl in Old New York	1896	1896	CP	1896		Dodd, Mead and Company
A Little Girl of Long Ago	1897	1897	CP/PREF	1897		A. L. Burt Company
Her Place in the World	1897	1897	CP/PREF	1897		Lee and Sheapard Publishers
Hannah Ann	1897	1897	CP/PREF	1897		Dodd, Mead & Company
A Little Girl in Old Boston	1898	1898	CP/PREF	1898		A. L. Burt Company
The Heir of Sherburne	1899	1899	PREF/CP	1899		Dodd, Mead & Company
A Little Girl in Old Washington	1900	1900	CP/PUB	1900	1900	Dodd, Mead & Company
A Little Girl in New Orleans	1901	1901	CP/PUB	1901	1901	Dodd, Mead & Company
A Little Girl in Old Detroit	1902	1902	CP/PUB	1902	1902	A. L. Burt Company
Helen Grant's Schooldays	1903	1903	CP/PUB	1903	1903	Lothrop, Lee & Shepard Co.
A Little Girl in Old St Louis	1903	1903	CP/PUB	1903	1903	Dodd, Mead & Company
How Bessy Kept House	1903	1903	CP	1903		
A Little Girl in Chicago	1904	1904	CP/PUB	1904	1904	Dodd, Mead & Company
A Little Girl in Old San Francisco	1905	1905	CP/PUB	1905	1905	Dodd, Mead and Company
Helen Grant at Aldred House	1905	1905	CP	1905	1906	Lee and Shepard
A Little Girl in Old Quebec	1906	1906	CP	1906		A. L. Burt Company
An Easter Lily	1906	1906	CP	1906		
Helen Grant Senior	1907	1907	CP/PUB	1907	1907	Lothrop, Lee & Shepard Co.
A Little Girl in Old Salem	1908	1908	CP/PUB	1908	1908	Dodd, Mead and Company
A Little Girl in Old Pittsburg	1909	1909	CP/PUB	1909	1909	A. L. Burt Company
Helen Grant Teacher	1909	1909	CP/PUB	1909	1909	Lothrop, Lee & Shepard Co
The Children in the Old Red House	1912?	1912	CP/PUB	1912	1912	Lothrop, Lee & Shepard Co
Modern Cinderella	1913	1913	CP	1913		M. A. Donohue & Company
The Red House Children at Grafton	1913	1913	CP/PUB	1913	1913	Lothrop, Lee & Shepard Co.
The Girls at Mount Morris	1914	1914	CP	1914		M. A. Donohue & Co.

Table A.5 – Collected works for Constance Fenimore Woolson.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
The Old Stone House	1873	1873	EAC	1873		O. Mothrop & Co.
The Ancient City	1873	1874	PUB		1874	Harper's New Monthly Magazine
Castle Nowhere	1865?	1875	CP/PUB	1875	1875	James R. Osgood and Company
Rodman the Keeper	1880	1880	CP	1880	1886	Harper Brothers
Anne	1880?	1882	EAC	1882		Harper & Brothers
For the Major	1883	1883	EAC/PUB	1883		Harper & Brothers
East Angels	1884	1884	CP	1884,1885,1886		Harper & Brothers
Jupiter Lights	1889	1889	CP/PUB	1889	1889	
Horace Chase	1894	1894	CP/PUB	1894	1894	Harper & Brothers
Dorothy	1892	1895	CP	1895	1896	
Metone, Cairo and Corfu†	1895	1895	CP	1895	1896	
The Front Yard and Other Italien Stories†	1895	1895	PUB		1895	

'†' indicates posthumously published works.

Table A.6 – Collected works for Marion Harland.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
Alone	1854	1854	EAC	1854	1856	J. C. Derby, 119 Nassau Street
Husks	1863	1863	EAC/PUB	1863	1863	Sheldon & Company
At Last	1870	1870	PUB		1870	
Jessamine	1873	1873	EAC	1873		G. W. Carleton & Co., Publishers
Breakfast, Luncheon and Tea	1875	1875	CP	1875	1886	Charles Scribner's Sons
Dinner Year-Book	1878	1878	CP	1878	1883	Charles Scribner's Sons
Common Sense in the Household: A Manual of Practical Housewifery^{R}	1871	1880	CP	1880	1883	Charles Scribner's Sons
Loitering in Pleasant Paths	1880	1880	CP/ PUB	1880	1880	Charles Scribner's Sons
Marion Harland's Cookery for Beginners	1884	1884	CP	1884,1893		D. Lothrop Company
Mr. Wayt's Wife's Sister	1894	1894	CP	1894		The Cassell Publishing Co.
The Secret of a Happy Home	1896	1896	CP	1896		The Christian Herald
When Grandmamma Was New	1899	1899	CP	1899		Lothrop Publishing Company
MHA: The Story of a Long Life	1910	1909	PREF			Harper & Brothers
Marion Harland's Complete Etiquette^{RC}	1914	1914	CP	1905,1907,1914		The Bobbs-Merrill Company

'*R*' indicates (substantial) revisions. '*C*' indicates collaborations.

Table A.7 – Collected works for Harriet Beecher Stowe.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
Uncle Tom's Cabin	1852	1852	PUB		1852	Beadbury and Evans, Printers Whitefriars
A Key to Uncle Tom's Cabin	1853	1853	EAC/PUB	1853	1853	John P. Jewett & Co.
Sunny Memories of Foreign Land	1854	1854	EAC/PUB	1854	1854	Sampson, and Company
Sunny Memories of Foreign Land Vol.2	1854	1854	EAC/PUB	1854	1854	Sampson, and Company
The May Flower and Miscellenaous Writings	1855	1855	PUB	1855	1855	Phillips, Sampson, and Company
The Minister's Wooing	1859	1859	PREF/PUB		1859	Phillips, Sampson, and Company
The Pearl of Orr's Island	1861	1862	CP	1862,1890,1896	1896	Houghton, Mifflin and Company
Agenes of Sorrento	1862	1862	CP	1862,1890,1896	1899?	Houghton, Mifflin & Co.
Men of Our Times	1868	1868	EAC/PUB	1868	1868	Hartford Publishing Co
The Byron Controversy	1870	1870	PUB		1870	Sampson Low, Son, and Marston
Pink and White Tyranny	1871	1871	EAC/PUB	1871	1871	Roberts Brothers
Oldtown Fireside Stories	1872	1871	EAC	1871	1872	Boston: James R. Osgood & Company
My Wife and I	1872	1871	EAC	1871	1872	J. B. Ford and Company
Palmetto-Leaves	1873	1873	EAC/PUB	1873	1873	James R. Osgood And Company
Women in Sacred History	1873	1873	EAC	1873	1874	J. B. Ford and Company
We and Our Neighbours	1875	1875	CP	1875		J. B. Ford & Company
Poganuc People	1878	1878	CP	1878		Fords, Howard, & Hulbert
The Salem Witchcraft...	1886	1886	PUB		1886	Fowler & Wells Co., Publishers

Table A.8 – Collected works for Elizabeth Stuart Phelps Ward.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
Gypsy Cousin's Joy	1866	1866	EAC	EAC:1866,CP:1895		Dodd, Mead and Company
Gypsy Breynton	1866	1866	EAC	EAC:1866,CP:1893		Dodd, Mead and Company
Gypsy Sowing and Reaping	1866	1866	EAC	EAC:1866,CP:1896		Dodd, Mead and Company
Gypsy's Year at the Golden Crescent	1867	1867	EAC	1867		Dodd, Mead and Company
The Gates Ajar	1868	1868	EAC	1868	1873	James R. Osgood and Company
Men, Women and Ghosts	1868	1869	EAC / PREF /PUB	1869	1869	University Press: Welch, Bigelow &., Cambridge
The Silent Partner	1871	1871	EAC/PUB	1871	1871	James R. Osgood and Company
Doctor Zay	1882	1882	CP/PUB	1882	1882	Houghton, Mifflin and Company
Beyond the Gates	1883	1883	CP	1883	1884	Houghton, Mifflin and Company
The Gates Between	1887	1887	PUB		1887	Ward, Lock and Co.,
Jack the Fisherman	1887	1887	CP	1887		Houghton, Mifflin Company
The Struggle for Immortality	1884–89	1889	CP	1889		Houghton, Mifflin and Company
Donald Marcy	1893	1893	CP	1893		Houghton, Mifflin and Company
A Singular Life	1894	1894	CP	1894	1896	Houghton, Mifflin and Company
Chapters from a Life	1894–96?	1896	CP	1896	1900	Houghton, Mifflin and Company
The Supply at Saint Agathas	1896	1896	CP	1896	1897	Houghton, Mifflin and Company
The Story of Jesus Christ	1897	1897	CP/PUB	1897	1897	Houghton, Mifflin and Company
Within the Gates	1900	1900	CP	1900	1901	Houghton, Mifflin and Company
Avery	1901	1901	CP	1901,1902	1902	Houghton, Mifflin and Company
Trixy	1904	1904	CP/PUB	1904	1904	Houghton, Mifflin and Company
Though Life Do Us Part	1907	1907	CP	1907, 1908	1908	Houghton, Mifflin and Company

Table A.9 – Collected works for Susan Warner.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
Wide, Wide World	1850?	1850	EAC	1850	1851	George P. Putnam, 155 Broadway
Wide, Wide World Vol.2	1850?	1850	EAC	1850	1851	George P. Putnam, 155 Broadway
Queechy	1852	1852	CP	1852	1854	Tauchnitz
Queechy Vol.2	1852	1852	CP	1852	1854	Tauchnitz
The Law and Testimony	1853	1853	EAC/PUB	1853	1853	Robert Carter & Brothers
Hills of the Shatemuc	1856	1856	PUB		1856	Tauchnitz
Say and Seal	1860	1860	PUB/PREF		1860	Tauchnitz
Say and Seal Vol.2	1860	1860	PUB/PREF		1860	Tauchnitz
The Children of Blackberry Hollow	1861	1861	EAC	1861		American Sunday-School Union
Melbourne House	1864	1864	EAC	1864	1865	Robert Carter & Brothers
Melbourne House Vol.2	1864	1864	EAC	1864	1865	Robert Carter & Brothers
The Old Helmet	1864	1864	PUB		1864	Tauchnitz
The Old Helmet Vol.2	1864	1864	PUB		1864	Tauchnitz
The Word: Walks from Eden	1866	1866	PUB		1866	James Nisbet
Daisy	1868?	1868	PUB		1868	Ward Lock Edition
Daisy in the Field	1868?	1868	PUB		1868	Ward Lock Edition Butler & Tanner Ltd
What She Could	1870?	1870	EAC	1870	1873	Robert Carter and Brothers
Opportunities	1871	1871	EAC/PUB	1871	1871	Robert Carter and Brothers
The House in Town	1871?	1871	EAC	1871	1872	Robert Carter and Brothers
Trading	1872	1872	EAC	1872	1873	Robert Carter and Brothers
The Gold of Chickaree	1876	1876	CP	1876		G. P. Putnam's Sons
Wych Hazel	1876	1876	CP	1876	1888	G. P. Putnam's Sons
Diana	1877?	1877	CP/PUB	1877	1877	G. P. Putnam's Sons
Pine Needles^{C}	1877	1877	EAC/PUB	1877	1877	Robert Carter & Brothers
My Desire	1879?	1879	CP/PREF	1879		Robert Carter & Brothers
The End of a Coil	1880	1880	CP/PREF	1880		Robert Carter & Brothers
A Letter of Credit	1881?	1881	CP/PREF	1881	1882	Robert Carter & Brothers
Nobody	1882	1882	PREF			James Nisbet & Co Limited
A Red Wallflower	1884	1884	CP/PREF/PUB	1884	1884	Robert Carter & Brothers

'*C*' indicates collaborations.

Table A.10 – Collected works for Edith Wharton.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
The Decoration of Houses^{C}	1897	1897	CP	1897		Charles Scribner's Sons
Italian Villas and Their Gardens	1904	1903	CP	1903,1904	1904	New York: The Century Co.
The Fruit of the Tree	1907	1907	CP	1907		Charles Scribner's Sons
Tales of Men and Ghosts	1910	1910	PUB		1910	
The Custom of the Country	1913	1913	PUB		1913	
Kerfol	1916	1916	CP	1916		Charles Scribner's Sons
Xingu	1911?	1916	CP	1916		Charles Scribner's Sons
Summer	1917	1917	PUB		1917	
The Marne	1918	1918	PUB		1918	Macmillan and Co., Limited
In Morocco	1920	1919	CP	1919,1920	1920	Charles Scribner's Sons

'*C*' indicates collaborations.

Table A.11 – Collected works for Horatio Alger jr.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
Helen Ford	1866	1866	EAC	1866		The John C. Winston Co.
Timothy Crump's Ward	1866	1866	PUB		1866	
Charlie Codman's Cruise	1866	1866	EAC	1866	1867	A. K. Loring, Publishers
Street Life in New York with the Boot-Blacks	1868	1868	PREF			
Fame and Fortune	1868	1868	EAC/PREF	1868		A. K. Loring, Publishers
Popular Juvenile Books	1869	1869	EAC/PREF	1869		A. K. Loring, Publishers
Mark the Match Boy	1869	1869	CP/PREF	1869		The John C. Winston Co.
Rufus and Rose	1870	1870	PREF			Philadelphia: Porter & Coates
Tattered Tom or The Story of a Street Arab	1871	1871	EAC/PREF	1871		A. K. Loring, Publishers
Phil the Fiddler	1872	1872	PREF			
Risen from the Ranks	1874	1874	PREF/PUB		1874	
The Young Outlaw or, Adrift in the Streets	1875	1875	EAC/PREF	1875		A. K. Loring Publishers
The Luggage Boy	1870	1876	PREF			The John C. Winston Co.
The Telegraph Boy	1879	1879	PREF			
The Young Miner	1879	1879	PREF/CP	1879		Henry T. Coates & Co.
From Canal Boy to President	1881	1881	PREF/PUB		1881	American Publishers Corporation
From Farm Boy to Senator	1882	1882	CP/PREF	1882		J. S. Ogilvie & Company
Ben's Nugget	1882	1882	PREF/CP	1882		The John C. Winston Co.
The Backwoods Boy	1883	1883	PREF/CP	1883		David Mckay, Publisher
Bob Burton	1888	1888	CP	1888		Porter & Coates
The Erie Train Boy	1890	1891	CP	1891		United States Book Company
Digging for Gold	1892	1892	CP	1892		Porter & Coates
Dan the Newsboy	1893	1893	CP	1893		A. L. Burt, Publisher
In a New World Among the Gold-Fields of Australia	1893	1893	CP	1893		Porter & Coates
The Disagreeable Woman	1895	1895	CP	1895		G. W. Dillingham, Publisher
Frank and Fearless	1897	1897	CP	1897		The John C. Winston Co.
A Boy's Fortune	1898	1898	CP	1898		The John C. Winston Co.
The Young Bank Manager	1898	1898	CP	1898		The John C. Winston Co.
Ruperts' Ambition	1899	1899	CP	1899		The John C. Winston Co.
Jed the Poorhouse Boy	1899	1899	CP	1899		The John C. Winston Co.
Mark Mason's Victory	1899	1899	CP	1899		A. L. Burt, Publisher
Adventures of a Telegraph Boy OR 'Number 91'	1899	1899	CP	1899,1900		H. M. Caldwell Company
Adrift in New York	1895	1900	PUB		1900	A. L. Burt Company, Publishers
A Debt of Honour†	1900	1900	CP	1900		A. L. Burt Company, Publishers
Bernard Brooks' Adventure†	1903	1903	CP	1903		A. L. Burt Company, Publishers (New York)
From Farm to Fortune†	1905?	1905	CP	1905		Grosset & Dunlap, Publishers
Randy of the River†	1906	1906	CP	1906		Grosset & Dunlap, Publishers

'†' indicates posthumously published works.

Table A.12 – Collected works for Timothy Shay Arthur.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
The Mother	1847	1847	PUB		1847	Henry Fanners
Lessons in Life, for All Who Will Read Them	1851	1851	PUB		1851	Philadelphia
The Lights and Shadows of Real Life	1851	1851	PUB		1851	Philadelphia
The Two Wives	1851	1851	PUB		1851	Philadelphia
Words for the Wise	1847	1851	PUB		1851	Philadelphia
Woman's Trials; or, Tales and Sketches from the Life Around Us	1851?	1851	PUB		1851	Philadelphia
True Riches or, Wealth Without Wings	1852	1852	EAC/PUB	1852	1852	Boston: L. P. Crown & Co., 61 Cornhill
Married Life: its Shadows and Sunshine	1852	1852	PUB		1852	Philadelphia
Heart-Histories and Life-Pictures	1852	1852	PREF		1853	New York
Finger Posts on the Way of Life	1853	1853	PUB		1853	Philadelphia
The Iron Rule; or, Tyranny in the Household	1853	1853	PUB		1853	Philadelphia
Home Lights and Shadows	1853	1853	PUB		1853	Philadelphia
Ten Nights in a Bar Room	1854	1854	EAC/PUB	1854	1854	Boston: L. P. Crown & Co., 61 Cornhill
The Good Time Coming	1855	1855	PUB		1855	Boston
The Wedding Guest	1856	1856	PUB		1856	Chicago, Ill.:
Friends and Neighbours	1856	1856	PUB		1856	Philadelphia
Words of Cheer for the Tempted, the Toiling, and the Sorrowing	1856	1856	PUB		1856	Philadelphia
The Hand but not the Heart; or, the Life-Trials of Jessie Loring	1859	1858	PUB		1858	New York
Lizzy Glenn	1859	1859	PUB		1859	Philadelphia
Trials and Confessions of a Housekeeper	1854	1859	PUB		1859	Philadelphia
The Allen House or, Twenty Years Ago and Now	1860	1860	PUB		1860	Philadelphia
Nothing But Money	1865	1865	EAC	1865		Carleto JY, Publisher, 413 Broadway
After a Shadow, and Other Stories	1868	1868	PUB		1868	New York
After the Storm	1868	1868	PUB		1868	Philadelphia
Three Years in a Man-Trap	1872	1872	EAC/PUB	1872	1872	Philadelphia J. M. Stoddart & Co.
Cast Adrift	1872	1873	PUB		1873	
Danger or; Wounded in the House of a Friend	1872	1875	PUB		1875	Philadelphia
The Seen and the Unseen	1877?	1877	EAC/PUB	1877	1877	Philadelphia: J. B. Lippincott & Co.
Hair Breath Escapes†	1889?	1889	CP	1889		Worthington Co., 747 Broadway
Adventures by Sea and Land†	1890	1890	PUB		1890	

'†' indicates posthumously published works.

Table A.13 – Collected works for Robert W. Chalmers.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
In the Quarter	1894	1894	CP	1894		F. Tennyson Neely, Publisher
The King in Yellow	1895	1895	PUB		1895	
The Mystery of Choice	1896–97	1896	PREF	1897	1897	New York: D. Appleton and Company
Lorraine	1897	1897	CP/PREF	1897		Harper & Brothers
The Adventures of a Modest Man	1900?	1900	CP		1900,1904,1905,1910,1911	Harper & Brothers
Cardigan	1901	1901	CP	1901		A. L. Burt Company, Publishers
The Maid at Arms	1902	1902	PUB/PREF		1902	
The Maids of Paradise	1902–03	1902	CP	1902	1903	Harper & Brothers, Publishers
In Search of the Unknown	1904?	1904	CP/PUB/PREF	1904?	1904	Harper & Brothers, Publishers
Special Messenger	1908–09	1904	CP	1904,1905,1908,1909	1909	D. Appleton and Company
The Reckoning	1905	1904	PREF	1905	1905	Braunworth & Co. Bookbinders and Printers Brooklyn, N.Y.
The Tracer of Lost Persons	1906	1906	PUB		1906	
The Fighting Chance	1906	1906	CP/PUB	1906	1906	Toronto Mcleod Allen, Publishers
The Younger Set	1907	1907	PUB		1907	D. Appleton and Company: New York
The Firing Line	1908	1908	PUB		1908	D. Appleton and Company: New York
The Danger Mark	1909	1909	PUB		1909	
The Green Mouse	1910	1910	PUB		1910	
Ailsa Paige	1910	1910	CP/PUB	1910	1910	D. Appleton and Company: New York and London
The Common Law	1911	1911	PUB		1911	New York and London: D. Appleton and Company
Japonette	1911–12	1911	CP	1911,1912	1912	D. Appleton and Company: New York and London
The Gay Rebellion	1911?	1911	CP	1911,1913		D. Appleton and Company: New York and London
The Streets of Ascalon	1912	1912	CP/PUB	1912	1912	D. Appleton and Company: New York and London
The Business of Life	1913?	1912	CP	1912,1913		D. Appleton and Company: New York and London
Quick Action	1913-14	1913	CP	1913,1914		D. Appleton and Company: New York and London
The Hidden Children	1914	1913	PREF		1914	
Athalie	1914–15	1914	CP	1914, 1915	1915	D. Appleton and Company: New York and London
Who Goes There!	1915?	1915	CP/PUB	1915	1915	D. Appleton and Company: New York and London
Police	1915	1915	PUB/PREF		1915	D. Appleton and Company: New York and London
The Girl Phillipa	1915	1915	CP	1915, 1916	1919	D. Appleton and Company: New York and London
Babarians	1917	1915	CP	1915, 1916. 1917	1917	A. L. Burt Company/D. Appleton and Company
The Dark Star	1916–17	1916	CP	1916, 1917	1917	A. L. Burt Company/D. Appleton and Company
The Restless Sex	1917–18	1917	CP	1917,1918		A. L. Burt Company/D. Appleton and Company
The Laughing Girl	1918	1918	CP	1918		A. L. Burt Company/D. Appleton and Company
The Moonlit Way	1918-19	1918	CP	1918, 1919		D. Appleton and Company
The Crimson Tide	1919	1919	CP	1919		A. L. Burt Company/D. Appleton and Company
The Slayer of Souls	1919–20	1919	CP	1919, 1920		New York: George H. Doran Company
The Little Red Foot	1920–21	1920	CP	1920, 1921		New York: George H. Doran Company
The Flaming Jewel	1922	1922	CP	1922		Triangle Books

Table A.14 – Collected works for Francis Marion Crawford.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
Mr Isaacs	1882	1882	PUB		1882	
Dr Claudius	1883	1883	PUB		1883	London: Macmillan and Co.
To Leeward	1883	1883	CP	1883		P. F. Collier & Son: New York
A Roman Singer	1893	1883	CP/PUB	1883,1884,189	1883,1894,1896,1898,1901,1906	London: Macmillan and Co.
An American Politician	1884	1884	PREF			
Marzio's Crucifix and Zoroaster	1885	1885	CP	1885,1887	1908	London: Macmillan and Co.
The Children of the King	1885	1885	PUB		1885	P. F. Collier & Son: New York
A Tale of a Lonely Place	1886	1885	PREF		1886	
Saracinesca	1887	1887	PUB		1887	
Paul Patoff	1887	1887	CP	1887,1892,1893,1894,1899,1906,1912	1911	London: Macmillan and Co.
Sant' Ilario	1888	1888	CP	1888	1889,1890,1891,1893,1895,1898,1901	Macmillan and Co., Limited
Greifenstein	1889	1889	PREF/PUB		1889	Macmillan and Co., Limited
A Cigarette-maker's Romance	1890	1890	CP	1890	1893	Macmillan and Co., Limited
Khaled: a Tale of Arabia	1891	1891	CP/PUB	1891	1891,1892	
Don Orsino	1891	1891	PUB		1891	New York: Grosset & Dunlap, Publishers
The Three Fates	1891	1891	CP	1891	1892,1893	Macmillan and Co.
Pietro Ghisleri	1892	1892	CP	1892	1893	Macmillan and Co.
Katherine Lauderdale	1892	1893	CP	1893	1894	Macmillan and Co.
Katherine Lauderdale Vol.2	1893?	1893	CP	1893	1894	Macmillan and Co.
Marion Darche	1893	1893	CP/PUB	1893	1893	Macmillan and Co.
The Ralstons	1893	1893	CP	1893	1894,1895,1899,1902	Macmillan and Co.
Wandering Ghosts	1894	1894	CP	1894,1899,1903,1905,1908,1911	1911	Macmillan and Co.
Adam Johnstone's Son	1895	1895	CP	1895, 1896, 1897		P. F. Collier & Son: New York
Taquisara	1895	1895	PUB		1895	
Corleone	1896	1896	CP	1896	1897,1898,1902,1905	Macmillan and Co.
A Rose of Yesterday	1897	1897	CP	1897	1897	Macmillan and Co.
Ave Roma Immortalis	1898	1898	CP	1898	1899	Macmillan and Co.
Ave Roma Immortalis Vol.2	1898	1898	CP	1898	1899	Macmillan and Co.
Via Crucis	1898	1898	CP	1898	1899	Macmillan and Co.
In the Palace of the King	1900	1900	PUB		1900	
Marieta: a Maid of Venice	1901?	1901	PUB		1901	P. F. Collier & Son: New York
Cecilia: a Story of Modern Rome	1902	1902	CP/PUB	1902	1902	Macmillan and Co.
The Heart of Rome	1903	1903	CP/PUB	1903	1903	Macmillan and Co.
Whosoever Shall Offend	1904	1905	PUB		1905	
Fair Margaret	1905	1905	CP/PUB	1905	1905(Reprinted:1906,1908,1909,1910)	New York: Grosset & Dunlap
Arethusa	1906–7	1906	CP	1906, 1907	1907	Macmillan and Co.
A Lady of Rome	1906	1906	CP/PUB	1906	1906	
The Diva's Ruby	1907	1907	CP	1907	1908	Macmillan and Co.
The Primadonna	1907	1908	PUB		1908	
The White Sister	1908	1908	CP	1908,1909	Reprinted:1909,1910,1911,1913	A. L. Burt Company Publishers: New York
Stradella	1908	1908	CP	1908,1909	1909	The Macmillan Company

Table A.15 – Collected works for Mark Twain.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
Innocents Abroad	1869	1869	PREF		1869	
Roughing It	1872	1972	EAC		1979	Hartford, Conn.: American Publishing Company
The Gilded Age	1873	1873	PUB			
Life on the Mississippi	1874	1874	CP	1874, 1875,1883,1899,1903		Harper & Brothers, Publishers
Old Times on the Mississippi	1876	1876	PUB			Toronto: Belford Brothers
The Adventures of Tom Sawyer	1875	1876	PREF			
A Tramp Abroad	1880	1880	PUB			
The Prince and the Pauper	1881	1881	CP	1881,1899,1909,1921		
The Adventures of Huckleberry Finn	1884	1884	CP	1884,1896,1899,1912		Harper & Brothers, Publishers
A Connecticut Yankee in King Arthur's Court	1889	1889	PREF			
The American Claimant	1892	1891	PREF		1892	
Those Extraordinary Twins	1894	1894	CP			
Tom Sawyer Abroad	1894	1894	PUB			Chatto & Windus, Piccadilly
The Tragedy of Pudd'nhead Wilson	1894	1893	CP	1894,1893–1894		Hartford, Conn. (Century Company)
Joan of Arc	1896	1896	CP			Chatto & Windus, Piccadilly
Following the Equator	1897	1897	PUB			Hartford: New York
A Double Barrelled Detective Story	1902	1902	CP		1902	Harper & Brothers
Chapters from My Autobiography	1907	1906	PREF		1907	North American Review
Chapters from My Autobiography	1907	1907	PREF		1907	North American Review
Christian Science	1907	1907	PREF			
Extract from Captain Stormfield's Visit to Heaven	1909	1909	CP			Harper & Brothers
Is Shakespeare Dead ?	1909	1909	PUB			Harper & Brothers, Publishers
The Mysterious Stranger†*	1916	1916	CP			Harper & Brothers, Publishers

'†' indicates posthumously published works. '*' indicates works that remained unfinished by the author.

Table A.16 – Collected works for Henry James.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
The American	1877	1877	PUB		1877	
The Europeans	1878	1878	CP		1879	Houghton, Osgood and Company
An International Episode	1878	1878	EAC	1878		Harper & Brothers, Publishers
Daisy Miller	1897	1897	PREF			
Hawthorne	1897	1897	PUB		1879	Macmillan and Co.
Confidence	1879	1880	CP		1880	Houghton, Mifflin and Company
Washington Square	1880	1880	EAC	1880	1901	Harper & Brothers, Publishers
The Portrait of a Lady	1881	1881	CP	1881	1882	Houghton, Mifflin and Company
Roderick Hudson	1879	1883	PUB		1883	Macmillan and Co.
A Little Tour of France	1884	1885	PUB		1885	Bernard Tauchnitz
Georgina's Reasons	1884	1885	PUB			
The Bostonians	1886	1886	CP		1886	Macmillan and Co.
Princess Casamassima	1886	1886	CP		1886	Macmillan and Co.
The Aspern Papers	1888	1888	PUB		1888	Macmillan and Co.
The Reverberator	1888	1888	CP		1888	Macmillan and Co.
The Tragic Muse	1890	1890	PUB		1891	Macmillan Company
Picture and Text	1893	1893	CP			Harper and Brothers
The Spoils of Poynton	1896	1896	CP		1897	Houghton, Mifflin and Company
The Other House	1896	1896	CP		1897	The Macmillan Company
What Maise Knew	1897	1897	CP		1897	Herbert S. Stone & Co.
The Two Magics	1898	1898	CP	1898	1898	Macmillan Company
In the Cage	1898	1898	CP	1898	1898	Herbert S. Stone & Company
The Awkward Age	1899	1899	CP		1899	Harper & Brothers, Publishers
The Sacred Fount	1901	1901	CP		1901	Charles Scribner's Sons
The Wings of the Dove	1902	1902	CP	1902	1902	Charles Scribner's Sons
The Wings of the Dove	1909	1902	CP	1902,1909	1909	Charles Scribner's Sons
The Ambassador	1903	1903	PUB		1903	Methuen & Co.
The Golden Bowl	1904	1904	PUB		1904	
Italien Hours	1909	1909	PUB		1909	
The Outcry	1911	1911	PUB		1911	
*The Ivory Tower**†	1917	1917	CP		1917	Charles Scribner's Sons
*The Sense of the Past**†	1917	1917	CP		1917	Charles Scribner's Sons

'†' indicates posthumously published works. '*' indicates works that remained unfinished by the author.

Table A.17 – Collected works for Harold McGrath.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
The Grey Cloak	1903	1903	PUB		1903	Grosset and Dunlap, Publishers
The Princess Elopes	1905	1905	PUB		1905	Grosset and Dunlap, Publishers
Hearts and Masks	1905	1905	PUB		1905	Grosset and Dunlap, Publishers
The Best Man	1907	1907	CP	1907		A. L. Burt Company, Publishers
The Lure of the Mask	1908	1908	CP	1908		The Bobbs-Merrill Company, Publishers
The Goose Girl	1909	1909	PUB		1909	Indianapolis The Bobbs-Merrill Company, Publishers
Splendid Hazard	1910	1910	CP	1910		Grosset & Dunlap
The Carpet from Bagdad	1911	1911	CP	1911		The Bobbs-Merrill Company, Publishers
The Place of Honeymoons	1912	1912	CP	1912		The Bobbs-Merrill Company, Publishers
Parrot & Co	1913	1913	CP	1913		A. L. Burt Company, Publishers
The Adventures of Kathlyn	1914	1914	CP	1914		The Bobbs-Merrill Company, Publishers
The Million-Dollar Mystery	1915	1915	CP	1915		Grosset & Dunlap
The Voice in the Fog	1915	1915	PUB		1915	Grosset & Dunlap
The Pagan Madonna	1920	1920	CP	1920,1921	1921	Garden City, N.Y., and Toronto Doubleday, Page & Company
The Ragged Edge	1922	1922	CP/PUB	1922	1922	Garden City, N.Y., and Toronto Doubleday, Page & Company

Table A.18 – Collected works for Edgar Saltus.

Title	First PUB Date	Assigned Dating Date	Source	CP Dates	PUB Dates	Publisher
Balzac	1884	1884	CP/PUB	1884	1884	Boston: Houghton, Mifflin and Company
The Philosophy of Disenchantment	1885	1885	CP	1885	1887	Boston: Houghton, Mifflin and Company
Mr. Incoul's Misadventure	1887	1887	CP	1887		Gilliss Brothers & Turnure
Eden: an Episode	1888	1888	CP/PREF	1888		Belford, Clarke & Company, Publishers
The Truth About Tristem Varick	1888	1888	CP/PREF	1888		Belford, Clarke & Company, Publishers
A Transient Guest, and Other Episodes	1889	1889	CP/PREF	1889		Belford, Clarke & Company, Publishers
The Pace That Kills	1889	1889	CP/PREF	1889		Belford, Clarke & Company, Publishers
Mary Magdalen	1891	1891	CP	1891		
Imperial Purple	1892	1892	CP	1892		
Enthralled	1894	1894	CP/PUB	1894	1894	The American News Company
Perfume of Eros	1905	1905	CP/PUB	1905	1905	A. Wessels Company
Historia Amoris	1906	1906	CP	1906		New York: Mitchell Kennerley
The Lords of the Ghostland	1907	1907	CP	1907		New York: Mitchell Kennerley
Daughters of the Rich	1909	1909	CP	1909		New York: The Macaulay Company
The Monster	1912	1913	PUB/CP	1913	1913	New York: Pulitzer Publishing Company
Oscar Wilde	1917	1917	CP/PUB	1917	1917	Chicago Brothers of The Book
The Paliser Case	1919	1919	PUB/CP	1919	1919	Boni and Liveright: New York

Table A.19 – Collected works for Upton Sinclair.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
A Prisoner of Morro	1898	1898	CP	1898		Street & Smith, Publishers
King Midas	1901	1901	PUB		1901	New York and London
A Cadet's Honour	1903	1903	CP	1903		Philadelphia: David Mckay
The Journal of Arthur Stirling	1903	1903	CP	1903		New York: D. Appleton and Company
On Guard or Mark Mallory's Celebration	1903	1903	CP	1903		Philadelphia: David Mckay
The Jungle	1906	1906	PUB		1906	
The Metropolis	1908	1908	PUB		1908	
The Moneychangers	1908	1908	PUB		1908	
Samuel the Seeker	1910	1910	PUB/CP	1910	1910	New York: B. W. Dodge & Company
Love's Pilgrimage	1911	1911	PUB		1911	Mitchell Kennerley: New York and London
Damaged Goods	1913	1913	CP	1913		The John C. Winston Company: Philadelphia
Sylvia's Marriage	1914	1914	CP	1914		the Author Long Beach, California
King Coal	1917	1917	PUB/CP	1917	1917	The Macmillan Company
Jimmy Higgins	1918	1918	CP	1918,1919	1919	Published By The Author Pasadena, Calif.
100% the Story of a Patriot	1920	1920	PUB		1920	Published By The Author Pasadena, Calif.
The Book of Life	1921	1921	CP	1921,1922		The Paine Book Company: Chicago
They Call Me Carpenter	1922	1922	PUB		1922	

Table A.20 – Collected works for William Dean Howells.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
Italian Journeys	1867	1867	EAC/PUB	1867	1867	Hurd and Houghton
Venetian Life^{R}	1866	1867	PREF			Hurd and Houghton
Surburban Sketches	1871	1871	EAC/PUB	1871	1871	
Their Wedding Journey	1871?	1871	PUB?			
A Chance Acquaintance	1873	1873	PUB		1873	Boston: James R. Osgood and Company
A Forgone Conclusion	1874	1874	EAC	1874	1875	Boston: James R. Osgood and Company
A Counterfeint Argument	1877	1877	CP/PUB	1877	1877	Boston: James R. Osgood and Company
The Lady of the Aroostook	1879	1879	CP/PUB	1879	1879	Boston: James R. Osgood and Company
A Modern Instance	1881	1881	CP	1881	1882	James R. Osgood and Company
Dr. Breen's Practice	1881	1881	PUB?		1881	
The Rise of Silas Lapham	1884	1884	CP	1884	1885	Boston: Ticknor and Company
Indian Summer	1885	1885	CP	1885	1886	Boston: Ticknor and Company
The Minister's Charge	1886	1886	CP	1886	1887	Boston: Ticknor and Company
April Hopes	1887	1887	PUB?		1887	
A Hazard of New Fortunes Vol.1	1889	1889	CP	1889	1890	New York: Harper & Brothers, Franklin Square
A Hazard of New Fortunes Vol.2	1889	1889	CP	1889	1890	New York: Harper & Brothers, Franklin Square
A Boy's Town	1890	1890	CP	1890		New York and London: Harper & Brothers, Publishers
The Quality of Mercy	1891	1892	PUB		1892	New York and London: Harper & Brothers, Publishers
The Coast of Bohemia	1893	1893	CP	1893,1899	1899	New York and London: Harper & Brothers, Publishers
A Traveller from Altruria	1894	1894	CP	1894	1908	
My Literary Passion	1895	1895	PUB		1895	
The Landlord at the Lion's Head	1896	1896	CP	1869,1897	1897	New York and London: Harper & Brothers, Publishers
Stories of Ohio	1897	1897	CP	1897		American Book Company
An Open-Eyed Conspiracy	1898	1898	PUB		1898	Edinburgh: David Douglas, Castle Street
A Ragged Lady	1899	1899	CP/PUB	1899	1899	New York and London: Harper & Brothers, Publishers
A Pair of Patient Lovers	1901	1901	PUB		1901	New York and London: Harper & Brothers, Publishers
The Kentons	1902	1902	PUB/CP	1902	1902	New York and London: Harper & Brothers, Publishers
The Flight of the Pony Baker	1902	1902	PUB/CP	1902	1902	New York and London: Harper & Brothers, Publishers
Questionable Shapes	1903	1903	PUB		1903	
Through the Eye of the Needle	1907	1907	PUB		1907	
Between the Dark and the Daylight	1907	1907	PUB		1907	
Roman Holiday and Others	1908	1908	CP/PUB	1908	1908	New York and London: Harper & Brothers, Publishers
Fennel and Rue	1908	1908	CP/PUB	1908	1908	New York and London: Harper & Brothers, Publishers
Seven English Cities	1909	1909	PUB/CP	1909	1909	New York and London: Harper & Brothers, Publishers
Imaginary Interviews	1910	1910	PUB/CP	1910	1910	New York and London: Harper & Brothers, Publishers
Familiar Spanish Travels	1913	1913	CP	1913		New York and London: Harper & Brothers, Publishers
The Daughter of the Storage	1915	1915	CP	1915,1916	1916	New York and London: Harper & Brothers, Publishers
Years of My Youth	1916	1916	CP	1916,1917	1917	New York and London: Harper & Brothers, Publishers

'*R*' indicates (substantial) revisions.

Table A.21 – Collected works for William Taylor Adams.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
All Aboard	1855	1855	PREF			Chicago: M.A. Donohue & Co.
Now or Never	1856	1856	EAC/PREF	1856		Boston: Brown, Bazin, and Company
Try Again	1857	1857	EAC	EAC:1857 CP:1885?		Boston: Lee and Shepard, Publishers
Poor and Proud	1858?	1858	PREF			
Little by Little	1860	1860	EAC/PREF	1860		Boston: Crosby, Nichols, Lee & Co.
In School and Out	1863	1863	EAC/PREF	1863		Boston: Lee and Shepard, Publishers
The Soldier Boy	1864?	1864	PREF			New York: Hurst & Company, Publishers
Watch and Wait	1864	1864	EAC/PREF	1864		Boston: Lee and Shepard, Publishers
Work and Win	1865	1865	PREF			New York: Hurst & Company, Publishers
The Young Lieutenant	1865	1865	EAC	1865		Boston: Lee and Shepard, Publishers
Hope and Have	1866	1866	EAC/PREF	1866		Boston: Lee and Shepard, Publishers
Brave Old Salt	1866	1866	EAC/PREF	1866		Boston: Lee and Shepard, Publishers
Haste and Waste	1866	1866	EAC/PREF	1866		Boston: Lee and Shepard, Publishers
Outward Bound	1866	1866	EAC/PREF	1866	1869	Boston: Lee and Shepard, Publishers
Breaking Away	1867	1867	EAC/PREF	EAC:1867 CP:1895		Boston: Lothrop, Lee & Shepard Co.
Seek and Find	1867	1867	EAC/PREF	EAC:1867 CP:1895		Boston: Lee and Shepard, Publishers
Down the River	1868?	1868	EAC/PREF	EAC:1868 CP:1896		Boston: Lee and Shepard, Publishers
Freaks of Fortune	1868	1868	EAC/PREF	EAC:1868 CP:1896		Boston: Lee and Shepard, Publishers
Make or Break	1868	1868	EAC/PREF	EAC:1868 CP:1896		Boston: Lee and Shepard, Publishers
Our Standard Bearer	1868	1868	EAC/PREF	1868		Boston: Lee and Shepard, Publishers
Dikes and Ditches	1868	1868	EAC/PREF	1868		Boston: Lee and Shepard, Publishers
Down the Rhine	1869	1869	EAC/PREF	1869		Boston: Lee and Shepard, Publishers
Field and Forest	1870	1870	EAC/PREF	1870		Boston: Lee and Shepard, Publishers
Desk and Debit	1870–71	1870	PREF	1871	1871	Boston: Lee and Shepard, Publishers
Plane and Plank	1870–71	1870	PREF	1871	1871	Boston: Lee and Shepard, Publishers
Up the Baltic	1871	1871	EAC/PUB	1871	1871	Boston: Lee and Shepard, Publishers
Northern Lands	1871	1871	PREF		1872	Boston: Lee and Shepard, Publishers
Little Bobtail	1872	1872	EAC	1872		Boston: Lee and Shepard, Publishers
The Yacht Club	1873	1873	EAC	1873		Boston: Lee and Shepard, Publishers
The Coming Wave	1874	1874	EAC	1874		Boston: Lee and Shepard, Publishers
Living Too Fast	1876	1876	EAC/CP/PREF	1876		Boston: Lee and Shepard, Publishers
Vine and Olive	1876	1876	CP/PREF	1876		Boston: Lee and Shepard, Publishers
Down South	1880	1880	CP/PREF	1880	1881	Boston: Lee and Shepard, Publishers
Up the River	1881	1881	CP/PREF	1881	1882	Boston: Lee and Shepard, Publishers
All Adrift	1882	1882	CP/PREF	1882	1883	Boston: Lee and Shepard, Publishers
Snug Habor	1883	1883	CP/PREF	1883	1884	Boston: Lee and Shepard, Publishers
Square and Compasses	1884	1884	CP/PREF	1884		Boston: Lee and Shepard, Publishers
Stem to Stern	1885	1885	CP/PREF	1885	1886	Boston: Lee and Shepard, Publishers
Within the Enemy's Lines	1889	1889	CP/PREF	1889	1890	Boston: Lee and Shepard, Publishers
On the Blockade	1890	1890	CP/PREF	1890		Boston: Lee and Shepard, Publishers
Stand by the Union	1891?	1891	CP/PREF	1891	1896	Boston: Lee and Shepard, Publishers
Fighting for the Right	1892	1892	CP/PREF	1892		Boston: Lee and Shepard, Publishers
A Victorious Union	1893	1893	CP/PREF	1893	1894	Boston: Lee and Shepard, Publishers
Asiatic Breezes	1894	1894	CP/PREF	1894	1895	Boston: Lee and Shepard, Publishers
In the Saddle	1894-95	1894	CP/PUB	1894	1895	Boston: Lee and Shepard, Publishers
Across India	1895	1895	PUB		1895	Boston: Lee and Shepard, Publishers
A Lieutenant at Eighteen	1895	1895			1895	Boston: Lee and Shepard, Publishers
The Boat Club^{R}	1870	1896	CP/PREF	1896		New York: The Mershon Company, Publishers
Four Young Explorers	1896	1896	CP/PUB	1896		Boston: Lee and Shepard, Publishers

'*R*' indicates (substantial) revisions.

Table A.22 – Collected works for Charles Dudley Warner.

Title	First PUB Date	Assigned Date	Dating Source	CP Dates	PUB Dates	Publisher
Saunterings	1872	1872	EAC/PUB	1872	1872	James R. Osgood and Company
Backlog Studies	1872	1872	EAC	1872	1885	Houghton, Mifflin and Company
Baddeck and That Sort of Thing	1874	1874	PREF			
In the Levant	1875?	1876	PUB/PREF		1876	Boston: Houghton, Mifflin and Company
Being a Boy	1877	1877	CP	1877,1897		Houghton, Mifflin and Company
How I killed a Bear	1878	1878	CP	1878,1888		Houghton, Mifflin and Company
My Winter on the Nile^{R}	1876	1880	PUB/PREF		1876	Boston: Houghton, Mifflin and Company
Captain John Smith	1881	1881	PREF			
Washington Irving	1881	1881	CP	1881	1884	Houghton, Mifflin and Company
Studies in the South and West with Comments on Canada	1889	1889	PUB		1889	New York: Harper & Brothers
Our Italy	1890?	1891	CP	1891		Harper & Brothers
The Golden House	1894	1894	CP	1894		B. W. Dodge & Company
The People for Whom Shakespeare Wrote	1897	1897	CP/PREF	1897		Harper & Brothers
That Fortune	1899	1899	CP	1899		Harper & Brothers, Publishers

'*R*' indicates (substantial) revisions.

Bibliography

Baayen, Harald; van Halteren, Hand, and Tweedie, Fiona. Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution. *Literary and Linguistic Computing*, 11(3):121–132, 1996. doi: 10.1093/llc/11.3.121. URL http://dx.doi.org/10.1093/llc/11.3.121.

Beach, Joseph Warren. *The Method of Henry James*. Yale University Press, 1918.

Blair, Walter. Reviewed Work: Twain and the Image of History by Roger B. Salomon. *American Literature*, 34(4):578–580, 1963. URL http://www.jstor.org/stable/2923090.

Branch, Edgar. Mark Twain and JD Salinger: A Study in Literary Continuity. *American Quarterly*, 9(2):144–158, 1957.

Burnham, Kenneth P. and Anderson, David R. Model Selection and Multimodel Inference: A Practical Information-theoretic Approach, 2003.

Burrows, John. 'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3):267–287, September 2002. URL http://dx.doi.org/10.1093/llc/17.3.267.

Burrows, John. All the Way Through: Testing for Authorship in Different Frequency Strata. *Language and Linguistic Computing*, 22(1):27–47, 2007.

Burrows, John F. Word-patterns and Story-shapes: The Statistical Analysis of Narrative Style. *Literary & Linguistic Computing*, 2(2):61–70, 1987.

Burrows, John F. 'An Ocean Where Each Kind...': Statistical Analysis and Some Major Determinants of Literary Style. *Computers and the Humanities*, 23(4-5):309–321, 1989.

Burrows, John F. Not Unless You Ask Nicely: The Interpretative Nexus between Analysis and Information. *Literary and Linguistic Computing*, 7:91–109, 1992.

Bybee, Joan. *Mechanisms of Change in Grammaticization: The Role of Frequency*, pages 602–623. Blackwell Publishing Ltd, 2008. ISBN 9780470756393. doi: 10.1002/9780470756393.ch19. URL http://dx.doi.org/10.1002/9780470756393.ch19.

Bybee, Joan and Thompson, Sandra. Three Frequency Effects in Syntax. In *Annual Meeting of the Berkeley Linguistics Society*, volume 23, 1997.

Can, Fazli and Patton, Jon M. Change of Writing Style with Time. *Computers and the Humanities*, 38(1):61–82, 2004.

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

Chaski, Carole E. Empirical Evaluations of Language-based Author Identification Techniques. *Forensic Linguistics*, 8:1–65, 2001.

Cheung, Hintat and Kemper, Susan. Competing Complexity Metrics and Adults' Production of Complex Sentences. *Applied Psycholinguistics*, 13(01):53–76, 1992.

Chong, Miranda and Specia, Lucia. Lexical Generalisation for Word-level Matching in Plagiarism Detection. In *International Conference Recent Advances in Natural Language Processing*, RANLP, pages 704–709, Hissar, Bulgaria, 2011. URL `http://clg.wlv.ac.uk/papers/ranlp-2011_chong.pdf`.

Chung, Cindy and Pennebaker, James W. The Psychological Functions of Function Words. *Social Communication*, pages 343–359, 2007.

Clark, Alexander Michael Simon. Forensic Stylometric Authorship Analysis under the Daubert Standard. *Journal of Law & Literature eJournal*, 2011.

Craig, Hugh. Authorial Attribution and Computational Stylistics: If You Can Tell Authors apart, Have You Learned Anything about Them? *Literary and Linguistic Computing*, 14(1): 103–113, 1999.

Daelemans, Walter. Explanation in Computational Stylometry. In *Computational Linguistics and Intelligent Text Processing*, pages 451–462. Springer, 2013.

Davies, Mark. 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*, pages 231–61, 2012.

Forsyth, Richard S. Stylochronometry with Substrings, or: A Poet Young and Old. *Literary and Linguistic Computing*, 14(4):467–478, 1999. doi: 10.1093/llc/14.4.467. URL `+http://dx.doi.org/10.1093/llc/14.4.467`.

Friedman, Jerome; Hastie, Trevor, and Tibshirani, Robert. *The Elements of Statistical Learning*, volume 1. Springer Series in Statistics Springer, Berlin, 2001.

Friedman, Jerome; Hastie, Trevor, and Tibshirani, Robert. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

Frontini, Francesca; Lynch, Gerard, and Vogel, Carl. Revisiting the Donation of Constantine. In *Proceedings of AISB*, pages 1–9, 2008.

Halteren, Hans and Oostdijk, Nelleke. Word Distributions in Dutch Tweets. A Quantitative Appraisal of the Distinction between Function and Content Words. 3:189–226, 01 2015.

Hamilton, William L; Leskovec, Jure, and Jurafsky, Dan. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

Henrich, Joseph; Heine, Steven J., and Norenzayan, Ara. The Weirdest People in the World? *Behavioral and Brain Sciences*, 33(2-3):61–83, 2010. doi: 10.1017/S0140525X0999152X.

Hoad, Timothy C and Zobel, Justin. Methods for Identifying Versioned and Plagiarized Documents. *Journal of the Association for Information Science and Technology*, 54(3):203–215, 2003.

Hoover, David L. Corpus Stylistics, Stylometry, and the Styles of Henry James. *Style*, 41(2): 174–203, 2007.

Hoover, David L. Quantitative Analysis and Literary Studies. *A Companion to Digital Literary Studies*, 2008.

Hornik, Kurt. *openNLP: Apache OpenNLP Tools Interface*, 2015. URL `https://CRAN.R-project.org/package=openNLP`. R package version 0.2-5.

Hornik, Kurt. *NLP: Natural Language Processing Infrastructure*, 2016. URL `https://CRAN.R-project.org/package=NLP`. R package version 0.1-9.

James, Gareth; Witten, Daniela; Hastie, Trevor, and Tibshirani, Robert. *An Introduction to Statistical Learning*, volume 112. Springer, 2013.

James, Henry. *The Art of Fiction*. Longmans, Green and Company, 1884.

James, Nicholas A and Matteson, David S. ecp: An R package for nonparametric multiple change point analysis of multivariate data. *arXiv preprint arXiv:1309.3295*, 2013.

James, Nicholas A; Kejariwal, Arun, and Matteson, David S. Leveraging Cloud Data to Mitigate User Experience from Breaking Bad. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 3499–3508. IEEE, 2016.

James D. Williams, . The Use of History in Mark Twain's A Connecticut Yankee. *PMLA*, 80 (1):102–110, 1965. URL http://www.jstor.org/stable/461131.

Jurafsky, Daniel; Bell, Alan; Gregory, Michelle, and Raymond, William D. Probabilistic Relations between Words: Evidence from Reduction in Lexical Production. *Typological Studies in Language*, 45:229–254, 2001.

Jurka, Timothy P.; Collingwood, Loren; Boydstun, Amber E.; Grossman, Emiliano, and van Atteveldt, Wouter. *RTextTools: Automatic Text Classification via Supervised Learning*, 2012. URL http://CRAN.R-project.org/package=RTextTools. R package version 1.3.9.

Kane, Michael J.; Emerson, John, and Weston, Stephen. Scalable Strategies for Computing with Massive Data. *Journal of Statistical Software*, 55(14):1–19, 2013. URL http://www.jstatsoft.org/v55/i14/.

Kemper, Susan; Greiner, Lydia H; Marquis, Janet G; Prenovost, Katherine, and Mitzner, Tracy L. Language Decline Across the Life Span: Findings from the Nun Study. *Psychology and Aging*, 16(2):227–239, 2001.

Kilgarriff, Adam. Comparing Corpora. *International Journal of Corpus Linguistics*, 6(1), 2001.

Kilgarriff, Adam. Language Is Never Ever Ever Random. *Corpus Linguistics and Linguistic Theory*, 1, 2005.

Klaussner, Carmen and Vogel, Carl. Stylochronometry: Timeline Prediction in Stylometric Analysis. In *Research and Development in Intelligent Systems XXXII*, pages 91–106. Springer, 2015.

Klaussner, Carmen and Vogel, Carl. Revisiting Hypotheses on Linguistic Ageing in Literary Careers. Paper presented at the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing) Budapest, Hungary, 2017.

Klaussner, Carmen and Vogel, Carl. A Diachronic Corpus for Literary Style Analysis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7–12 2018a. European Language Resources Association (ELRA).

Klaussner, Carmen and Vogel, Carl. Temporal Predictive Regression Models for Linguistic Style Analysis. *Journal of Language Modeling*, 6(1):175–222, 2018b.

Klaussner, Carmen; Vogel, Carl, and Bhattacharya, Arnab. Detecting Linguistic Change Based on Word Co-occurrence Patterns. In *HistoInformatics@ CIKM*, pages 14–21, 2017.

Koppel, Moshe; Argamon, Shlomo, and Shimoni, Anat Rachel. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002.

Koppel, Moshe; Schler, Jonathan, and Bonchek-Dokow, Elisheva. Measuring Differentiability: Unmasking Pseudonymous Authors. *Journal of Machine Learning Resources*, 8:1261–1276, December 2007. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1314498.1314541`.

Koppel, Moshe; Schler, Jonathan, and Argamon, Shlomo. Authorship Attribution in the Wild. *Language Resource Evaluation*, 45(1):83–94, March 2011. doi: 10.1007/s10579-009-9111-2. URL `http://dx.doi.org/10.1007/s10579-009-9111-2`.

Koppel, Moshe; Schler, Jonathan, and Argamon, Shlomo. Authorship Attribution: What's Easy and What's Hard. *Journal of Law and Policy*, 21:317, 2012.

Kotsakos, Dimitrios; Lappas, Theodoros; Kotzias, Dimitrios; Gunopulos, Dimitrios; Kanhabua, Nattiya, and Nørvåg, Kjetil. A Burstiness-aware Approach for Document Dating. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1003–1006. ACM, 2014.

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

Kuhn, Max. *Caret: Classification and Regression Training*, 2014. URL `http://CRAN.R-project.org/package=caret`. With contributions from: Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer and the R Core Team, R package version 6.0-30.

Le, Xuan; Lancashire, Ian; Hirst, Graeme, and Jokel, Regina. 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, 2011.

Lieberman, Erez; Michel, Jean-Baptiste; Jackson, Joe; Tang, Tina, and Nowak, Martin A. Quantifying the Evolutionary Dynamics of Language. *Nature*, 449(7163):713, 2007.

Lyse, Gunn Inger and Andersen, Gisle. Collocations and Statistical Analysis of N-grams. *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian, Studies in Corpus Linguistics, John Benjamins Publishing, Amsterdam*, pages 79–109, 2012.

Makridakis, Spyros; Wheelwright, Steven C, and Hyndman, Rob J. *Forecasting Methods and Applications*. John Wiley & Sons, 2008.

Manning, Christopher D. and Schütze, Hinrich. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-13360-1.

Mendenhall, Thomas Corwin. The Characteristic Curves of Composition. *Science*, 9 (214S):237–246, 1887. doi: 10.1126/science.ns-9.214S.237. URL http://www.sciencemag.org/content/ns-9/214S/237.short.

Mendenhall, Thomas Corwin. A Mechanical Solution of a Literary Problem. *The Popular Science Monthly*, 60:97–105, 1901.

Michalke, Meik. *koRpus: An R Package for Text Analysis*, 2014. URL http://reaktanz.de/?c=hacking&s=koRpus. (Version 0.05-4).

Milborrow, Stephen. *plotmo: Plot a Model's Response and Residuals*, 2017. URL https://CRAN.R-project.org/package=plotmo. R package version 3.3.4.

Milne, Alan Alexander and Shepard, Ernest Howard. *Winnie-the-Pooh*. Egmont, 2013.

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

Narayanan, Arvind; Paskov, Hristo; Gong, Neil Zhenqiang; Bethencourt, John; Stefanov, Emil; Shin, Eui Chul Richard, and Song, Dawn. On the Feasibility of Internet-Scale Author Identification. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, SP '12, pages 300–314, Washington, DC, USA, 2012. IEEE Computer Society. ISBN 978-0-7695-4681-0. doi: 10.1109/SP.2012.46. URL http://dx.doi.org/10.1109/SP.2012.46.

Newman, Matthew L; Pennebaker, James W; Berry, Diane S, and Richards, Jane M. Lying Words: Predicting Deception from Linguistic Styles. *Personality and Social Psychology Bulletin*, 29(5):665–675, 2003.

Niculae, Vlad; Zampieri, Marcos; Dinu, Liviu P, and Ciobanu, Alina Maria. Temporal Text Ranking and Automatic Dating of Texts. In *EACL*, pages 17–21, 2014.

Oakes, Michael P. *Literary Detective Work on the Computer*. John Benjamins Publishing Company, 2014. ISBN 9027249997, 9789027249999.

Pena, Edsel A. and Slate, Elizabeth H. *gvlma: Global Validation of Linear Models Assumptions*, 2014. URL http://CRAN.R-project.org/package=gvlma. R package version 1.0.0.2, (last verified: 24.08.2015).

Pennebaker, James W. The Secret Life of Pronouns: How Our Words Reflect Who We Are. *New York: Bloomsbury*, 2011.

Pennebaker, James W and Stone, Lori D. Words of Wisdom: Language Use Over the Life Span. *Journal of Personality and Social Psychology*, 85(2):291–231, 2003.

Pennebaker, James W; Francis, Martha E, and Booth, Roger J. Linguistic Inquiry and Word Count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(1), 2001.

Pinheiro, Jose; Bates, Douglas; DebRoy, Saikat; Sarkar, Deepayan, and R Core Team, . *nlme: Linear and Nonlinear Mixed Effects Models*, 2013. R package version 3.1-113.

R Core Team, . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.r-project.org`.

Rangel, Francisco and Rosso, Paolo. Use of Language and Author Profiling: Identification of Gender and Age. *Natural Language Processing and Cognitive Science*, 177, 2013.

Rangel, Francisco; Rosso, Paolo; Verhoeven, Ben; Daelemans, Walter; Potthast, Martin, and Stein, Benno. Overview of the 4th Author Profiling Task at PAN 2016: Cross-genre Evaluations. *Working Notes Papers of the CLEF*, 2016.

Revolution Analytics, and Weston, Steve. *foreach: Foreach Looping Construct for R*, 2014. URL `http://CRAN.R-project.org/package=foreach`. R package version 1.4.2.

Rudman, Joseph. Cherry Picking in Nontraditional Authorship Attribution Studies. *Chance*, 16(2):26–32, 2003.

Rybicki, Jan and Eder, Maciej. Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words? *Literary and Linguistic Computing*, 26(3):315, 2011. doi: 10.1093/llc/fqr031. URL `+http://dx.doi.org/10.1093/llc/fqr031`.

Schilleman, Matthew. Typewriter Psyche: Henry James's Mechanical Mind. *Journal of Modern Literature*, 36(3):14–30, 2013.

Schler, Jonathan; Koppel, Moshe; Argamon, Shlomo, and Pennebaker, James W. Effects of Age and Gender on Blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205, 2006.

Schmid, Helmut. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, volume 12, pages 44–49. Manchester, UK, 1994.

Smith, David L. The Jamesian Oedipus and the Freudian Moses: Image, Word, The Later Style, and The Ambassadors. *Studies in the Novel*, 44(1):1–26, 2012.

Smith, Joseph A. and Kelly, Colleen. Stylistic Constancy and Change across Literary Corpora: Using Measures of Lexical Richness to Date Works. *Computers and the Humanities*, 36(4): 411–430, 2002. URL http://www.jstor.org/stable/30204686.

Snowdon, David A. Healthy Aging and Dementia: Findings from the Nun Study. *Annals of Internal Medicine*, 139(5_Part_2):450–454, 2003.

Štajner, Sanja and Mitkov, Ruslan. 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 Cultural Heritage at RANLP*, pages 78–85, 2011.

Stamatatos, Efstathios. Ensemble-based Author Identification Using Character N-grams. In *In Proceedings of the 3rd Int. Workshop on Textbased Information Retrieval*, pages 41–46, 2006.

Stamatatos, Efstathios. A Survey of Modern Authorship Attribution Methods. *Journal of the Association for Information Science and Technology*, 60(3):538–556, March 2009. ISSN 1532-2882. doi: 10.1002/asi.v60:3. URL http://dx.doi.org/10.1002/asi.v60:3.

Stamou, Constantina. Stylochronometry: Stylistic Development, Sequence of Composition, and Relative Dating. *Literary and Linguistic Computing*, 23(2):181–199, 2007.

Tabata, Tomoji. Dickens's Narrative Style: A Statistical Approach to Chronological Variation. *Revue Informatique et Statistique dans les Sciences humaines (RISSH)*, 30:165–182, 1994.

Tweedie, Fiona J and Baayen, R Harald. How Variable May a Constant Be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, 32(5):323–352, 1998.

Venables, W. N. and Ripley, B. D. *Modern Applied Statistics with S*. Springer, New York, 4 edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.

Walsh, Thomas M. and Zlatic, Thomas D. Mark Twain and the Art of Memory. *American Literature*, 53(2):214–231, 1981. URL http://www.jstor.org/stable/2926100.

Watson, G. S. Inference and Disputed Authorship: The Federalist by Frederick Mosteller; David L. Wallace. *The Annals of Mathematical Statistics*, 37(1):pp. 308–312, 1966. ISSN 00034851. URL http://www.jstor.org/stable/2238718.

Wecter, Dixon. Mark Twain and the West. *Huntington Library Quarterly*, 8(4):pp. 359–377, 1945. ISSN 00187895. URL http://www.jstor.org/stable/3816065.

Zhao, Ying and Zobel, Justin. Searching with Style: Authorship Attribution in Classic Literature. In *Proceedings of the Thirtieth Australasian Conference on Computer Science - Volume 62*, ACSC '07, pages 59–68, Darlinghurst, Australia, 2007. Australian Computer Society, Inc. ISBN 1-920-68243-0. URL http://dl.acm.org/citation.cfm?id=1273749.1273757.

Zipf, George Kingsley. *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press, 1932.

Zou, Hui and Hastie, Trevor. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005. URL http://www.jstor.org/stable/3647580.