Diachronic Delta: A computational and dialectical method for analysing literary corpora

Chris Beausang

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Department of Computer Science
Faculty of Science and Engineering
Maynooth University

July 2021

Head of Department and Supervisor

Dr. Joseph Timoney

Contents

List of Tables

List of Figures

Abstract

Much has been written on computational literary studies' (CLS) rigidity and reductiveness in comparison with literary criticism's more pragmatic and intuitive means of approaching its object of study. This thesis attempts to undermine this antinomy via dialectical materialist philosophy as proposed by George Wilhelm Friedrich Hegel and developed by Karl Marx. As a science and a method of social critique, dialectics not only has a long history of usage within literary criticism, but they also provide a means of mediating the distinctions between empirical evidence, logic and intuition.

We do so by operationalising John Burrows' 'Delta' method across time rather than as it is conventionally applied, across text (J. Burrows, "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship"). We thereby identify particular years as being associated with extensive amounts of 'novelty' (which years introduce the most amount of distance from proceeding years) and also possess extensive amounts of 'resonance' (are relatively proximate to their successor years). This is undertaken on the basis that agents within a dataset which score highly for both of these metrics are highly innovative, as they are significantly different from the years which come before, and relatively similar to the years which come after (Barron et al., "Individuals, institutions, and innovation in the debates of the French Revolution."; Barron et al., "Supplementary Information: Individuals, Institutions, and Innovation in the Debates of the French Revolution"). We refer to these years as 'breaks' and identify them in order to contrast their behaviour with the *longue-durée* approach increasingly prevalent within CLS (Underwood, *Distant Horizons: Digital Evidence and Literary Change* 25).

Though this theory of incremental change ultimately remains in overall terms robust, this thesis nevertheless presents significant results arising from this method and demonstrates the ways in which dialectical materialist philosophy may ground the use of empirical methods more coherently within contemporary literary critical practice.

Acknowledgements

Thanks are first of all due to Dr. Martin Charlton, Dr. John Keating, Professor Emer Nolan and Dr. Joseph Timoney without whom it would be highly unlikely that this research project would ever have reached its terminal stages.

Thanks of no secondary order are also due to everyone I have worked with, both in and outside of the Arts and Humanities Institute over the past four years. For their patience, insight, feedback, technical assistance, book or film recommendations, proofreading and opportunities to blow off steam, as well as everyone in the Occasional Student and Graduate Workers' Union and the Marxist reading group, have all contributed so much. Not only in terms of work done, but also in making the last four years bearable.

I am tremendously grateful to those in the English department who provided me with advice and support in conducting both my teaching and research, such as Oona Frawley, Sinéad Kennedy and Conor McCarthy.

Last but by no means least, thank you to Helen, to whom this thesis is dedicated, with love.

Publications

The author wishes it to be known that they have been successful in placing a reasonable proportion of the material which appears below in peer-reviewed journals.

An article derived in large part from material outlined in this thesis' first chapter has appeared in *magazén* and an article giving an overview of this thesis' primary output from a methodological perspective, as outlined in this thesis' second and third chapters, has been recommended for publication subject to moderate changes in the paper's content in *Digital Scholarship in the Humanities.* At time of writing it is awaiting the editor-in-chief's final decision.

Beausang, Chris. A Brief History of the Theory and Practice of Computational Literary Criticism (1963-2020). *magazén*, 1(2):181-201, 2020. doi: 10.30687/mag/2724-3923/2020/02/002

Glossary

As this thesis straddles the boundary between literary criticism and statistical analysis, it is inevitable that some terms will be deployed which may be unfamiliar to specialists on one side of the disciplinary boundary as opposed to another. This section will therefore provide a reference point for some of these terms.

It should be noted that many of these terms which appear below, particularly those in the technical terms section, refer to very specific and well-defined methods or approaches which have a history of being operationalised within the context of mathematics or statistics. Those more native to literary criticism are more discursive and have, perhaps uniquely in the context of this thesis, been made to bear the weight of a specific definition in order to facilitate their being used as benchmarks in a quantitative analysis. It is therefore necessary to consider these terms from the point of view of their application as distinct from their meaning. Finally it should be noted that as scholars working within computational literary studies seek to draw from statistics with a view to answering what are, by computational standards, often quite impressionistic research questions, it is inevitable that their usage will increasingly come to be defined discursively. The attempt is made in what follows to do justice to both the empirical and philological aspects of each of these terms where appropriate.

Analysis of variance (ANOVA): Analyses of variance or ANOVA is a statistical method applied to categorical data. When there is a single factor with three or more levels, one-way ANOVA is applied. When there are two or more factors, two or three-way ANOVA is used. ANOVA is, somewhat paradoxically, premised on the comparing of means, via the analysis of the data's variance (Crawley 150–52; Smith and Kelly).

Bootstrap Consensus Trees: The advantages associated with applying bootstrapping methods to the distances between texts can be seen in *Figs.* 1 - 3. Each of these plots visualise the textual distances which exist between thirty-six novels written by eight different authors. As can be seen in *Fig.* 1, when the distances between these texts are calculated on the basis of the 900 MFWs, three texts written by Louisa May Alcott, *Pauline's Passion* (1862), *Behind a Mask* (1866) and *Flower Fables* (1854), are identified as being most proximate to the novels of Anne, Charlotte and Emily Brontë. However, when the distances between these texts are calculated on the basis of the 1000 MFWs, as is the case in *Fig.* 2, these texts are more proximate to the rest of Alcott's *oeuvre*. By aggregating the distances which arise across ten dendrograms, from 100 - 1000 MFWs, as is the case in 3 (Eder, "Visualization in stylometry: Cluster analysis using networks" 56), we get an overall view and see the result which arises in *Fig.* 2 represents the exception and these three works by Alcott are indeed far more similar to the Brontë's novels that they are to Alcott's other works. **See also: Distance, Most frequent words, n-grams**

Break: A break is the term this thesis uses in order to define a moment in cultural history which departs to a significant extent from all prior moments. Though this term cannot in itself be said to have a wide circulation within the history of literary criticism, its terms are implicit in the way many historical literary movements are critically regarded, for example in the way new criticism or Marxist literary criticism regards literary modernism of the early twentieth century.

Chi-square test: This is a test applied data outlining how many times a particular event was observed. The test seeks to identify the presence of a statistically significant difference between the frequencies which are expected and the frequencies which are observed (Lantz 87). **See also: Fisher test, t-test**

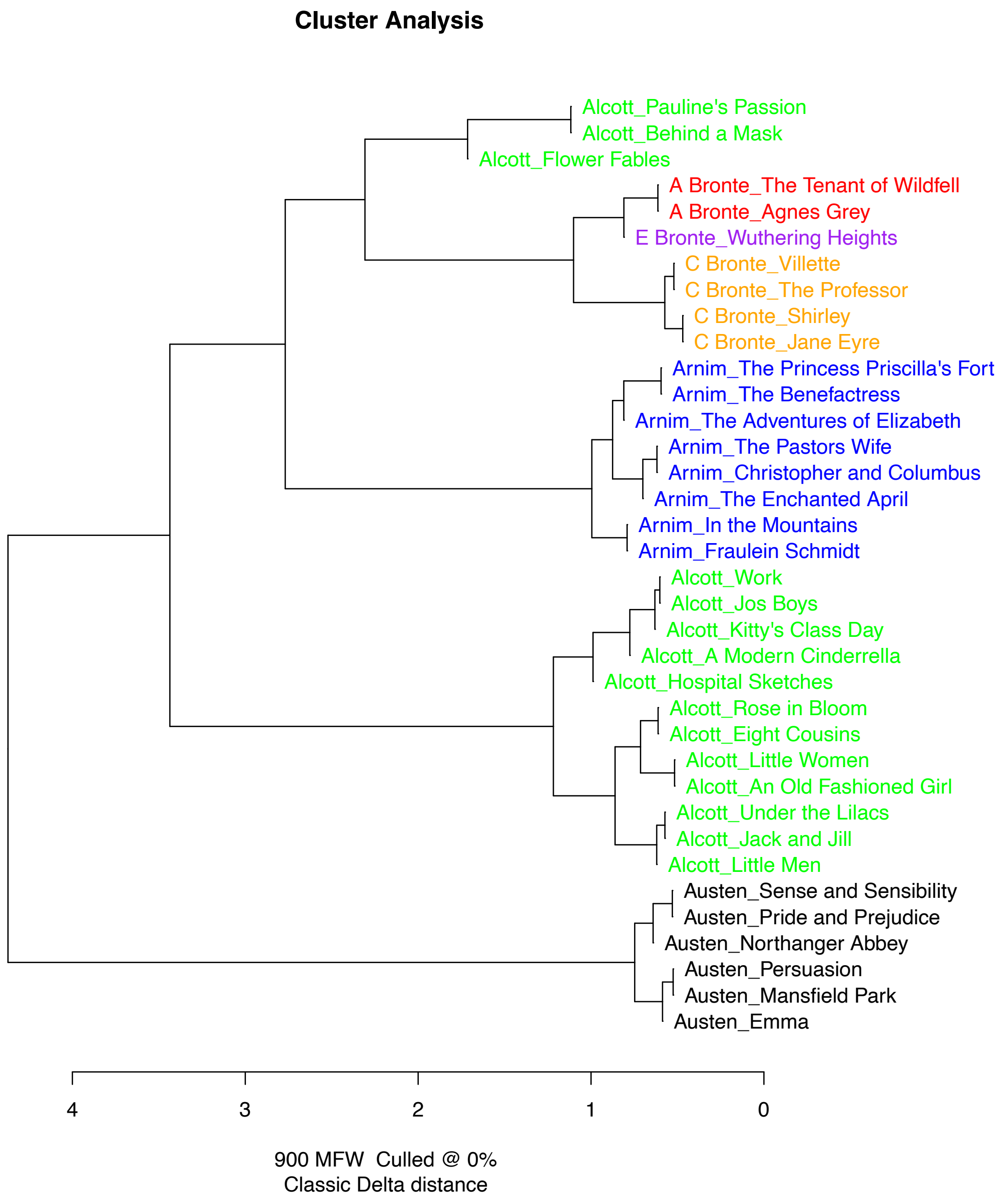


Figure 1: Dendrogram visualising distances based on 900 MFWs in 36 novels

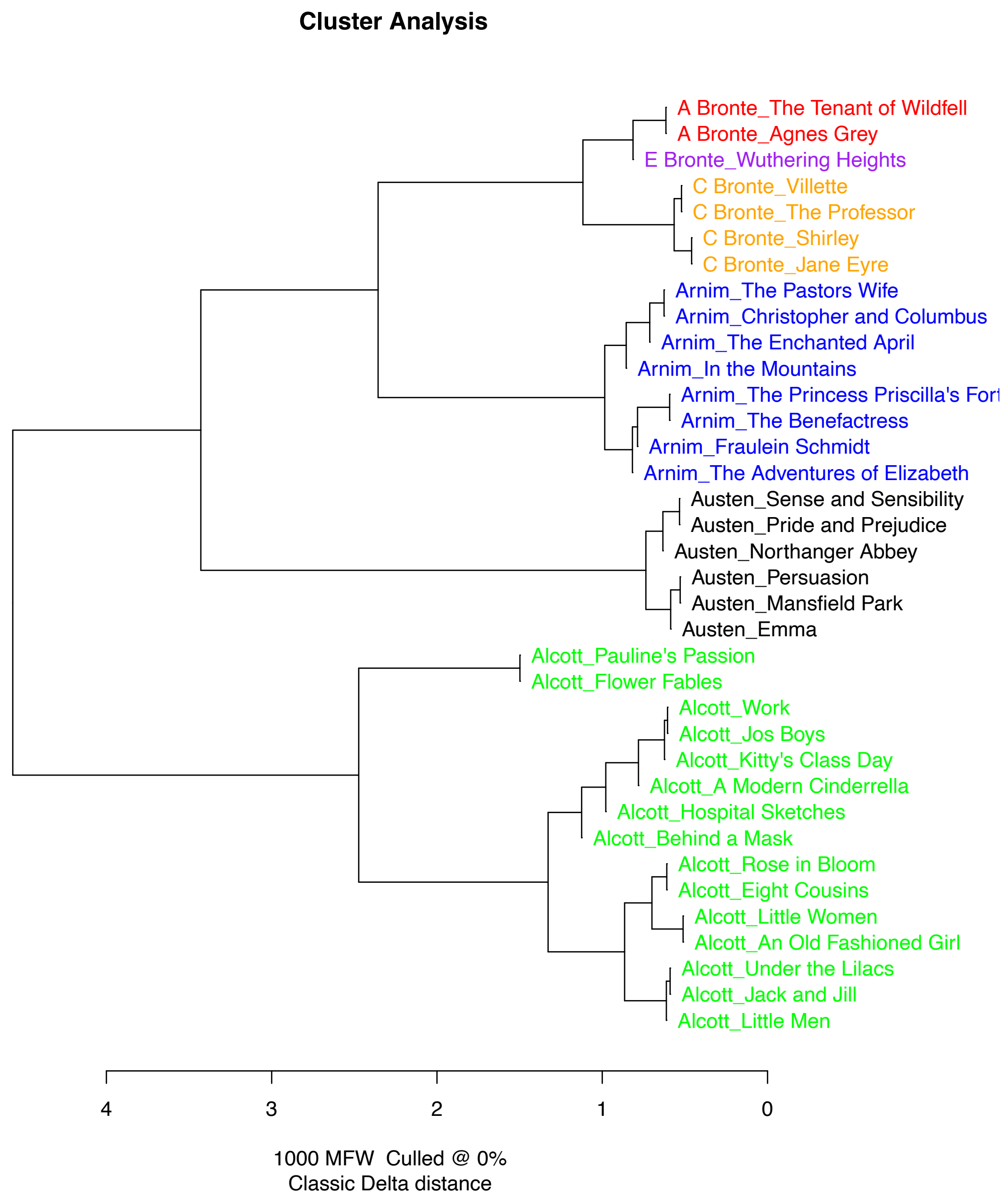


Figure 2: Dendrogram visualising distances based on 1000 MFWs in 36 novels

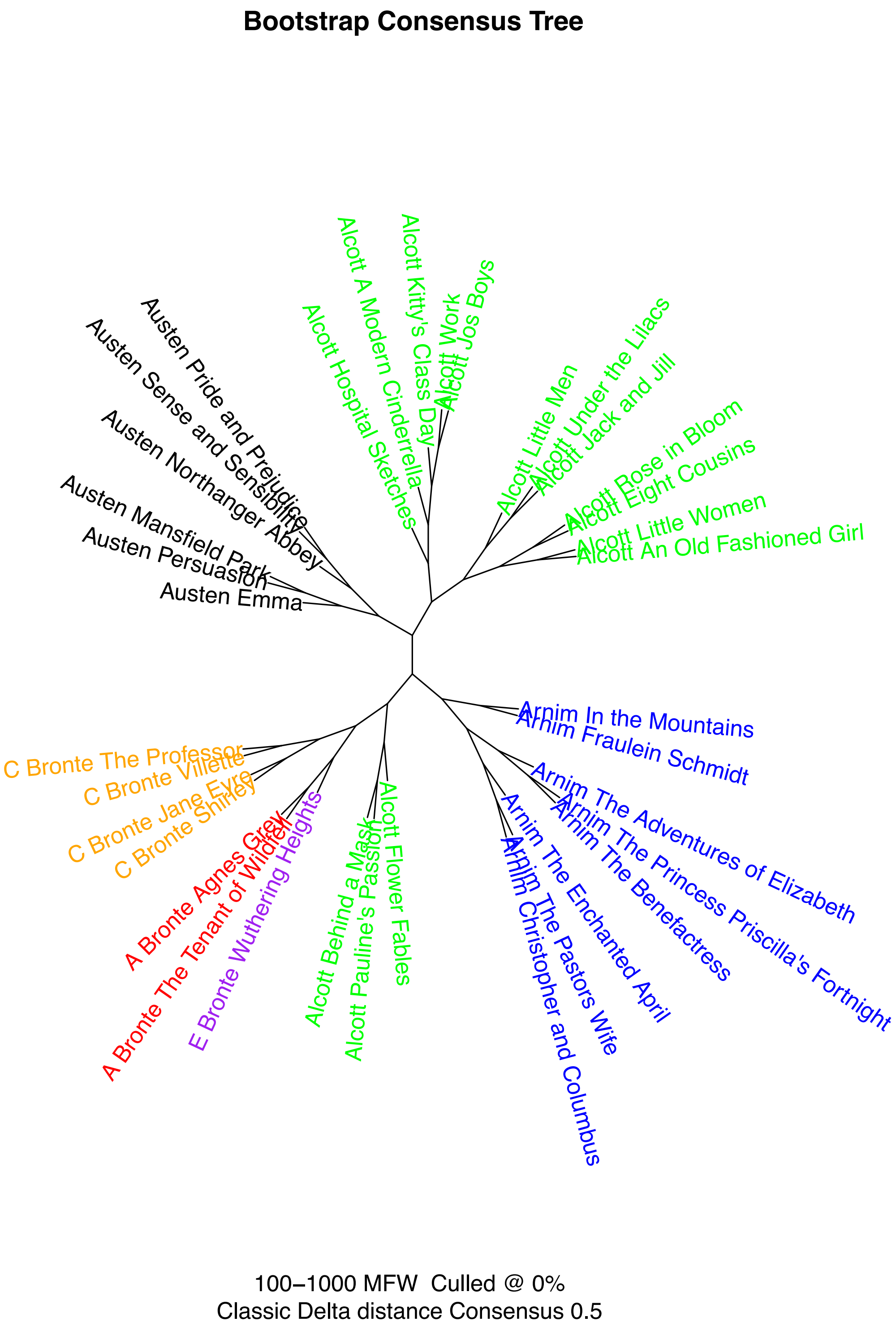


Figure 3: Bootstrap consensus tree visualising distances based on 1000 MFWs in 36 novels

Content words: This term refers to word-types which may be evaluated outside of their specific context, such as 'misgivings,' 'bewilderment' and 'account.' Where exactly content words arise in the frequency rank of a given corpus is difficult to pinpoint exactly, this will depend to a significant extent on the corpus under consideration, but in the three corpora this thesis analyses, they begin to arise most strongly after the 150th most frequent word in a corpus. **See also: Function words, Most Frequent Words**

Correlation co-efficients: A correlation co-efficient is a figure between -1 and 1 that is calculated between two variances. A value closer to 1 means greater similarity while closer to -1 means an inverse similarity (Crawley 108–10). Whether or not a correlation coefficient is statistically significant traditionally depends on their being either greater than or equal to 0.7 or less than or equal to -0.7.

Distance: In a stylometric context, distance refers to the difference between specific variables in a text and how this difference is calculated (J. Burrows, "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship"; Evert et al.). As implied by the word itself, a greater distance means a greater difference. There are as many ways to calculate distance as there are distance metrics. A fully worked example of how cosine distance, a variation on Burrows' Delta which has been shown to yield the most successful rates of authorship attribution in the context of a number of benchmark studies, (Evert et al.; Jannidis et al.) appears in the first Appendix to this thesis. **See also: Most Frequent Words**

Elastic net: A regularised logistic regression method which aims to combine the best of both ridge and LASSO regularisation within one algorithm. Elastic net combines a ridge regression into the penalty term in order to make the LASSO shrinkage more robust against the pitfalls usually associated with LASSO (Çiftsüren and Akkol 281; Lever et al.; Ogutu et al. 3)**See also: regression, ridge regression, least absolute shrinkage and selection operator regression (LASSO)**

Fisher test: This is a test applied data outlining how many times a particular event was observed. It seeks to identify the presence of a statistically significant difference between expected and observed frequencies, particularly in instances where the sample sizes are small (Crawley 105). **See also: Chi-square test, t-test**

Function words: This term refers to approximately 150 highly frequent word-types that are computed from a body of text. These consist of prepositions, conjunctions and articles, such as 'the,' 'an' and 'as' as tjese wprds serve a more overtly functional role purpose in the text (Holmes 112). **See also: Content words**

K-Nearest Neighbours (k-NN): *k*-Nearest-Neighbours is a well-known clustering technique used to automatically divide a dataset into a smaller number of smaller sets which have shared similarity. It is a distance-based measurement in the sense that in or out group membership will be determined based on the proximity of these data points in *k*-dimensional space (Temple 68; Zanin et al. 10).

Kullback-Leibler divergence: An entropy-based measurement used in order to assess the difference between two statistical populations that allows the identification of information which is lost in adopting one explanatory model as opposed to another (Kullback and Leibler). Kullback-Leibler divergence is often applied in stylometry to more general ends, as a means of computing the distance between two vectors (Pasanek and Sculley 354–55).

Least absolute shrinkage and selection operator regression (LASSO): LASSO is a regularised regression method, in that it suppresses the value of particular variables. What differentiates LASSO from other regularisation methods such as ridge, is that it is capable of reducing particular variables to zero, effectively removing them from the model altogether. LASSO therefore functions as a means of variable selection to a greater extent than ridge regression. This is one of the potential drawbacks of LASSO, in that the variables which are removed from the model are often arbitrary (Waldron et al. 3399). **See also: elastic net, regression, ridge regression**

Latent Dirichlet Allocation (LDA): A statistical method used in the context of stylometry to identify latent patterns in large text-based datasets and to construct what are referred to as topic models. The idea underpinning the construction and analysis of topic models is that sets of terms which occur together to a statistically significant extent form clusters of words which may be referred to as topics (Garcia-Zorita and Pacios 530; Viola and Verheul 6) and that it is possible to develop an understanding of a corpus' content from a macroscopic point of view by analysing these models.

LOESS: LOESS, or locally-weighted scatterplot smoothing, is a tool used in the visualisation of regression lines in order to render relationships between variables or trend lines more coherent (Upton and Cook).

Logistic regression: The aim of logistic regression is effectively the same as regression proper in that it aims to construct an explanatory model of a particular dataset with minimal amounts of error. Logistic regression is more suited to instances in which the response variable is a binary one, i.e. in instances where we are attempting to classify particular data points into one of two classes.

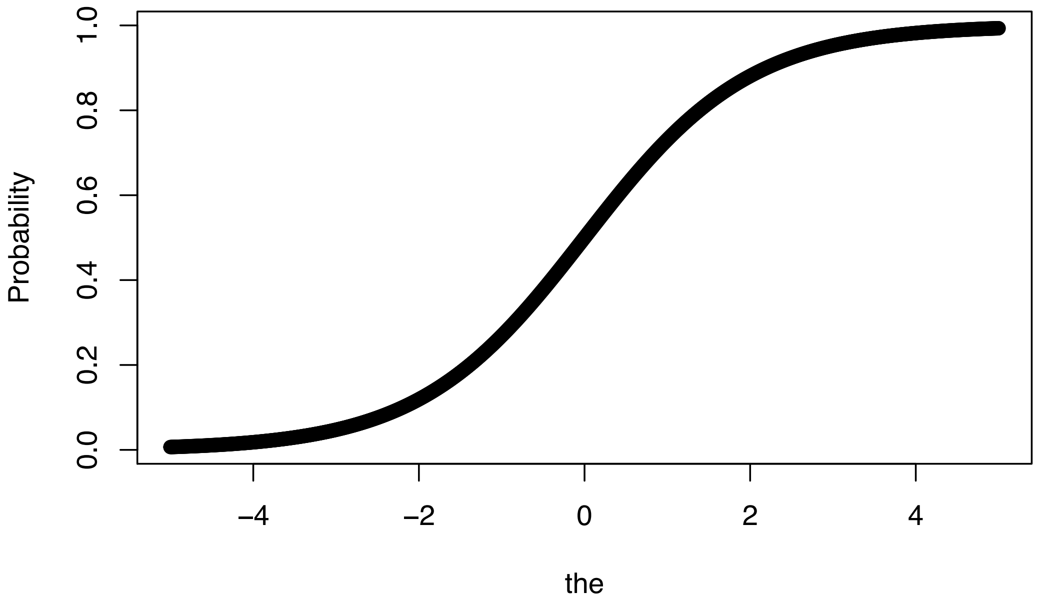


Figure 4: A sigmoid curve

We may see an example of how logistic regression works in practice in *Fig.* 4. This model represents an attempt to predict the certainty that a particular novel was written in the nineteenth century by modelling probability against normalised relative frequencies of the word 'the.' If a novel was to fall at zero along the *y*-axis, it would indicate absolute certainty that the text was not written in the nineteenth century, but if it were to align with one, it would indicate absolute certainty. In general we may say that increasing use of the word 'the' indicates successively greater amounts of confidence in a text being written in the nineteenth century. The utility of logistic regression may be seen from this example; if we were to attempt to develop a predictive model of this kind on a linear line, we would very quickly report certainties lower than zero or higher than one, a sigmoid line maintains our values within the limits of actual probabilities (McDonald 99; Underwood, *Distant Horizons: Digital Evidence and Literary Change* 194). **See also: elastic net, ridge regression, least absolute shrinkage and selection operator regression, regression**

Machine learning: As a term, 'machine learning' can refer to a wide range of methods and applications implemented using software on a computer, from prediction, automation, clustering, classification, or general data mining. Machine learning has come to be used widely in stylometry for the reason that it performs well in instances where we possess a wide range of variables and do not wish to impose an *a priori* model or method (Lantz 31). Machine learning therefore allows for the data, as Danilo Bzdok puts it, to 'speak for itself' (Bzdok et al.). **See also: Nearest Shrunken Centroids, Support Vector Machines**

Mann-Whitney test: The Mann-Whitney is a method of assessing whether or not there is a statistically significant difference in the median between two populations. Its focus on the median, as opposed to the arithmetic mean, renders it a non-parametric test (J. F. Burrows, "Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information" 97). The benchmark of statistical significance generally arises from the calculation of a *p*-value which is seemed to be so if it is less than 0.05 (McKenna and

Antonia 65). Unlike *t*-tests, Mann-Whitney tests do not assume the distributions of the two vectors are normal, i.e. follows the classic bell-shaped curve (Lijffijt et al. 380). **See also: t-test**

Modal auxiliaries: A modal auxiliary accompanies and modifies a verb, as in 'can,' 'may,' 'must' (J. F. Burrows, "Modal Verbs and Moral Principles: An Aspect of Jane Austen's Style" 1)

Most frequent words (MFWs): As the name suggests, most frequent words are the word-types which most frequently occur within a particular corpus or text are therefore used in computational literary studies as analytical units precisely for this reason. They are computed simply by finding the cumulative sum of instances of each word in a particular text and then assessing their overall relative frequency. Exactly how many MFWs are required in order for a stylometric analysis to be viable remains a contentious point within the literature, depending as it does on the objective of a study, text length as well as other factors but the tendency over time has been towards the inclusion of more MFWs rather than less, and thousands if possible (Rybicki and Eder 315–21).

***n*-grams:** Grams are units of text which provide the parameters for any given study. While single MFWs, or unigrams ('he,' 'she,' 'said') are the most commonly quantified units, bigrams may also be quantified ('he said,' 'she said') or trigrams ('they were saying,' 'we were saying'). More successful results have also been obtained through grams based on particular spans of characters as opposed to words. For example, five character grams we might extract from the phrase 'Stately plump Buck Mulligan,' of increasing size from one to five would include 'S,' 'St,' 'ly,' 'uck,' 'Mulli').

Naive-Bayes classification: A classification method where the probability of a vector belonging to a particular class is estimated. It should be noted htat Naive-Bayes classifiers assume that the value of a particular feature/vector in performing as a descriptor of the data is independent of the value of any other feature. (Hoorn et al. 316; Zanin et al. 9).

Nearest Shrunken Centroids (NSC): Another machine learning method used in the context of stylometry. NSC involves the calculation of a representative average for each class presented to the algorithm. In and out membership is then arrived at on the basis of the critical distance between this computed centroid and each observation (Schöberlein 645–46).

A noteworthy instance of NSC being applied is in the context of rolling stylometry. Rolling stylometry begins by the analyst supplying the algorithm with training data and a set number of classes. The performance of the model is then based on how successfully the algorithm can correctly assign a window of text of fixed length to each class. Shifting class membership from one window of text to another would suggest sudden shifts in writing style, signalling either precisely this or potential instances of collaborative authorship (Eder, "Rolling stylometry"; Ilsemann, "Stylometry approaching Parnassus").

A diagram illustrating how this works in practice can be seen in *Fig.* 5 which applies NSC analysis to *The Inheritors* (1901), a novel written collaboratively by both Ford Madox Ford and Joseph Conrad. Our training corpus consists of 21 prose texts written solely by Ford and Conrad. An inspection of the results we obtain would suggest that Ford is responsible for drafting the overwhelming majority of the novel but that there are not insignificant portions of roughly five thousand words each where Conrad's input was also significant. **See also: Machine learning, Support Vector Machines**

Principal Component Analysis (PCA): It is often the case in CLS that we may have a wide array of MFWs, into the thousands or even tens of thousands. However, when performing an analysis we may only be interested in a small set of these features which can capture the majority of the data. To reduce the number of features, a technique called Principal Components Analysis (PCA) may be employed. Applying the PCA technique means that those features which account for the greatest variance of the data are identified and retained, whereas those that have little influene are removed (Binongo and Smith; Joliffe 1). It should be noted that PCA assumes that all features are independent of one

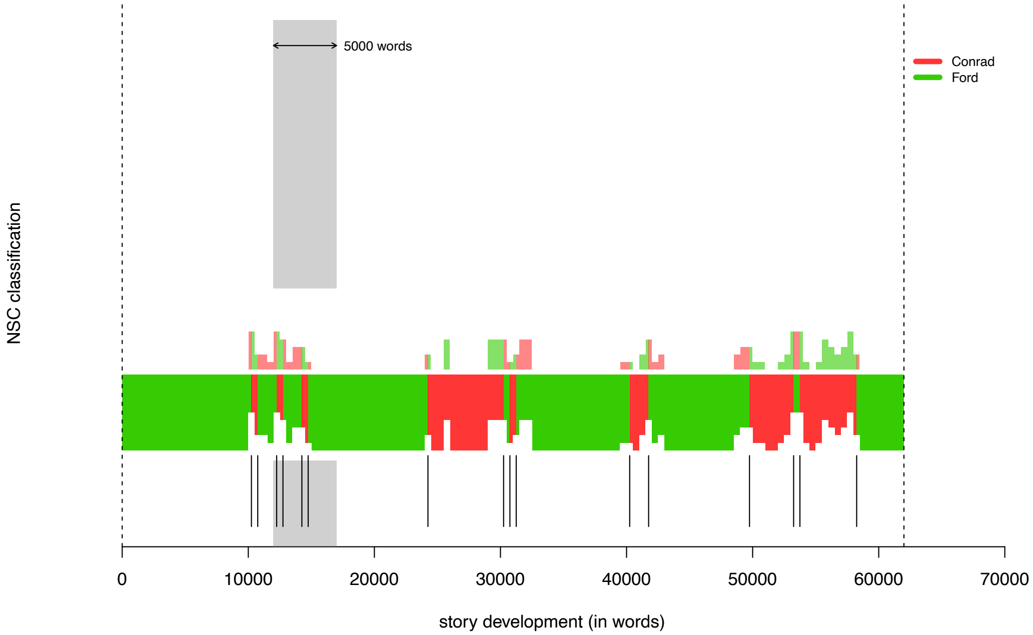


Figure 5: NSC classification of The Inheritors by Ford Madox Ford and Joseph Conrad

another.

Quartile: In accounting for the spread of the value of a particular variable, statisticians are usually attentive to the two most extreme values, the minimum and the maximum observed values. An example of which can be seen in *Fig.* 6.

The median accounts for the data's central tendency and is represented by the dividing line drawn within the box. The whisker closest to the x-axis represents the minimum value, while the whisker furthest from the x-axis is the maximum value. The first quartile refers to the value below which one quarter of the values are found and is represented by the lowermost part of the box. The third quartile refers to the value above which one quarter of the values are found, represented by the uppermost part of the box (Lantz 71–72).

Regression: Regression is a statistical technique used in order to model the relationship between variables, most commonly the relationship between a response variable on the

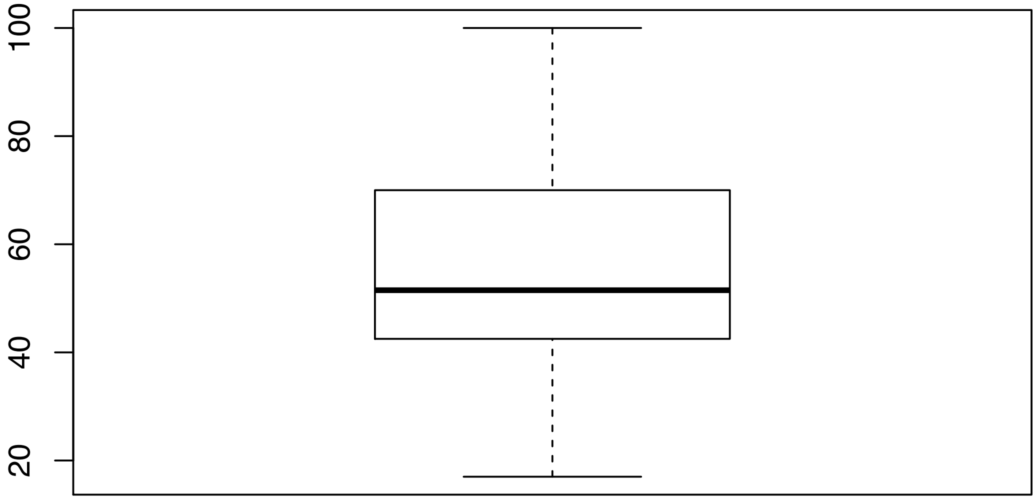


Figure 6: An example of quartile distributions on a boxplot

y-axis and an explanatory variable on the x-axis (Crawley 114; Moisl 135).

The distance between a predicted value and the observed value is called a residual and it is the aim of regression to minimise the value of these combined residuals over the whole input dataset multiplied by itself, a figure known as the sum of the squared residuals (SSR). This is the principle underpinning ordinary least squares regression (OLS), an example of which can be seen in *Fig.* 7.

In the line which bisects *Fig.* 7, we can see an attempt to model the relationship existing between two randomly generated variables, time on the x-axis and activity on the y-axis. In a number of instances we can see our model was reasonably successful and encounters a few of our data points, or does not miss them by much. In a number of instances though it has fallen quite wide of the mark and lines have been drawn between our regression line and the actual data points themselves; these represent our residuals. **See also: logistic regression, regularised regression**

Regularised regression: The aim of regularised regression is effectively the same as regression proper in that it aims to construct an explanatory model with minimal error. The difference between regularised regression and regression proper is that regularisation

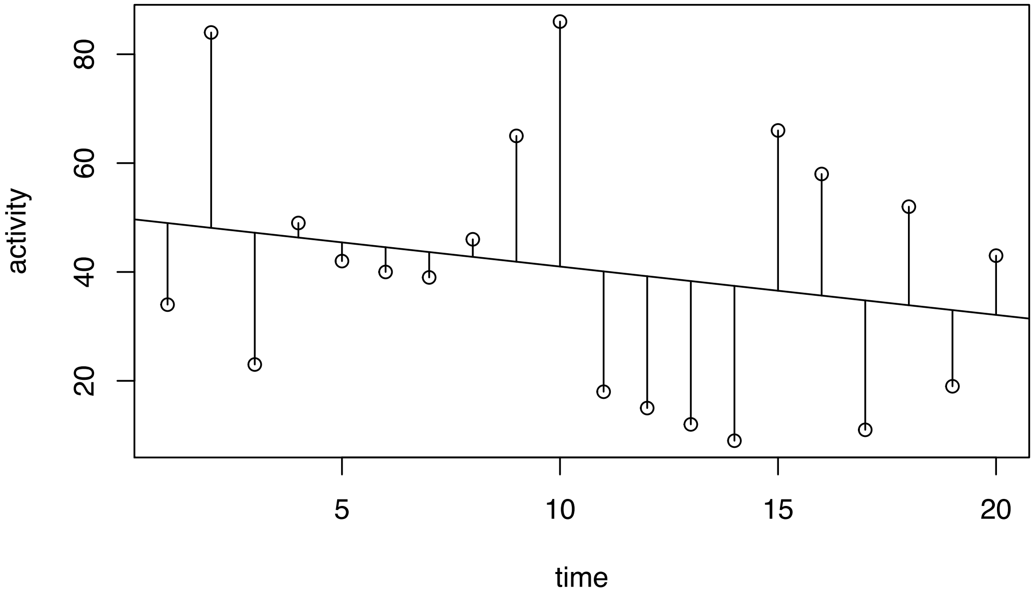


Figure 7: An example of ordinary least squares regression (OLS)

involves the calculation of a penalty term, referred to as lambda, which expresses the degree to which the value of particular variables will be reduced in order to prevent overfitting, which can result in the model not generalising well to the remainder of the data (McDonald 99). **See also: elastic net, ridge regression, least absolute shrinkage and selection operator regression**

Ridge regression: Ridge regression is a regularised regression method which shrinks the co-efficients of particular predictors towards zero, but does not remove them from the model altogether (Ogutu et al. 2). In this sense ridge regression is distinct from regularisation methods such as elastic net and LASSO which does remove particular co-efficients from the model. **See also: elastic net, regression, least absolute shrinkage and selection operator regression (LASSO)**

Support vector machines (SVM): A form of machine learning used within stylometry. SVM is based on the fitting of a linear boundary which facilitates the best separation between two or more classes of data, in much the same way that regression aims to arrive at a 'best fit' for a sequence of observations. The support vectors referred to in the method's name are the observations from each class which lie closest to this boundary and therefore play a significant role in determining the means through which the classification itself operates (Lantz 257–60).

A diagram indicating how SVM may be applied to stylometry can be seen in *Fig.* 8. This is a visualisation of an SVM model which has been trained on the major prose works of James Joyce; *Dubliners* (1914), *A Portrait of the Artist as a Young Man* (1916), *Ulysses* (1922) and *Finnegans Wake* (1939), as well as Samuel Beckett's *Molloy* (1951), *Malone Dies* (1951), *The Unnamable* (1953) and *Texts for Nothing* (1959). The model has been trained on seven random samples of 16000 words each as these eight novels alone would not be sufficient to train a sophisticated model. Though SVM can, and often is, applied to a wider range of variables, it is far easier to represent the results of a binary problem in two dimensions. In this instance, our model is only attentive to the relative usage of the words 'he' or 'she' in these random samples.

There is a significant amount of visual information to unpack in *Fig.* 8. It is first important to note that the clear section of the graph across the bottom represents the sector of the graph where the model places samples which approximate Beckett's writing style (1) to a greater extent. Beckett's writings seem to score lower in terms of usage of the terms 'he' (*y*-axis) and 'she' (*x*-axis), while the cross-hatched section of the graph along the top right, which score higher for both terms, represent samples the classifier understands as having been written by Joyce (2). The crosses, which are more proximate to the dividing line between these two sectors than the circles, represent the support vector machines, or those samples which play a significant role in determining where the classifier draws the dividing line.

We can identify what novels these vector machines are most often drawn from as can be seen in Table 1.

The circles represent the other samples which are seemingly less important in determining the position of this dividing line. The darker circles and crosses indicate Beckett's samples,

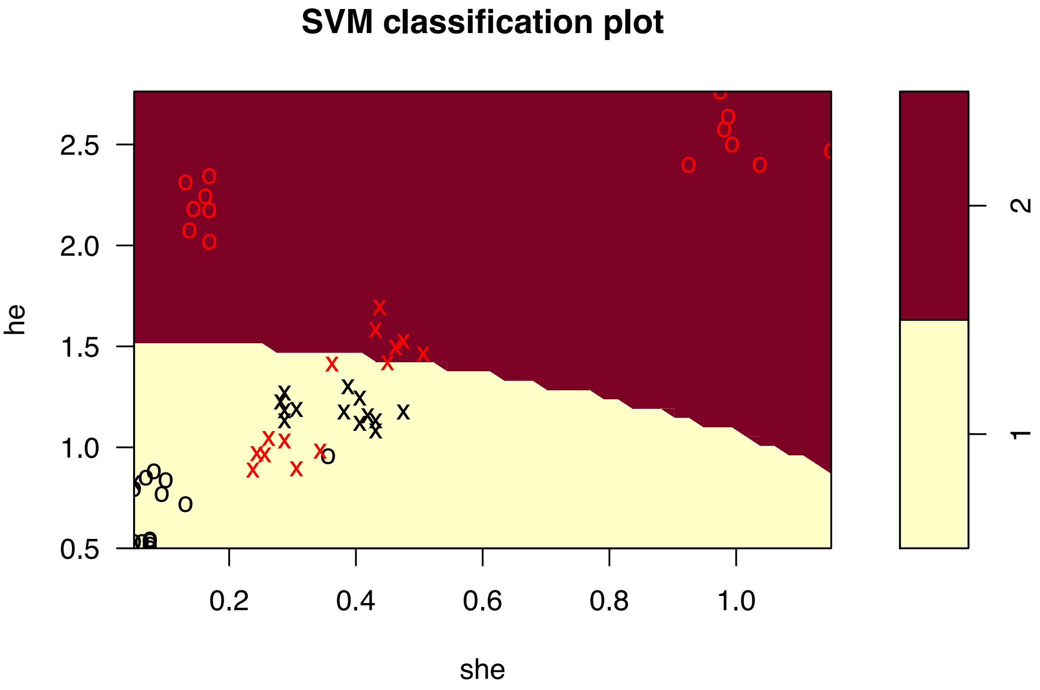


Figure 8: SVM classification of samples from the major prose works of Joyce and Beckett on the basis of their relative usage of gendered pronouns

Table 1: Most influential novels

Novels	Frequency
Beckett_Malone Dies	7
Joyce_Finnegans Wake	7
Joyce_Ulysses	7
Beckett_Molloy	6
Beckett_The Unnamable	1

while the lighter ones indicate Joyce. There seems to be a greater degree of spread for the Joyce data points in general, suggestive of Joyce's tendency to use gendered pronouns in a more heterogenous way than Beckett. **See also: Machine learning, Nearest Shrunken Centroids.**

t-test: A *t*-test is a method of assessing whether or not there is a statistically significant difference in the means between two populations. This focus on the arithmetic mean renders the t-test a parametric test. The benchmark of statistical significance generally arises from the calculation of a *p*-value which is less than 0.05. The *t*-test assumes the distributions are normally distributed, i.e. the histogram of the data exhibits a bell-shaped curve. This assumption can be unrealistic for certain types of real-world data which are not normally distributed, in which case a Mann-Whitney test would be preferred (Crawley 10; Lijffijt et al. 380). **See also: Mann-Whitney test**

z-scores: *z*-scores are the result of a standardisation method used in applications of Burrows' Delta method as well and many of its subsequent improvements. *z*-scores express each value in a vector of data points in terms of how many standard deviations it resides from a vector's mean, which is expressed as being equal to zero (J. Burrows, "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship" 271; Evert et al. ii6). The *z*-score is also known as the standard score. This is because is allows for the comparison of scores over different variables by standardising the distribution. Without this method of normalisation being applied to our relative word frequencies, only a corpus' most frequent words would play a decisive influence in the calculation of Delta, as these would be the significantly larger values in any given text.

Zeta: Zeta is a method proposed by Burrows in 2007 which aimed to bring a greater degree of attention to medium-frequency content words which are capable of differentiating one author from another (J. Burrows, "All the Way Through: Testing for Authorship in Different Frequency Strata" 27–28). For each word-type to be found in either corpus *a* or corpus *b*, a figure is calculated. This figure is equal to the number of segments in corpus *a* with one or more instances of this word, divided by the total number of segments in corpus *a*, plus the number of segments in corpus *b* with no instances, divided by the total number of segments in text *b*. The word-types with the highest index are those which are most useful in differentiating corpus *a* from corpus *b* (Rizvi).

Introduction

> One of the most important tasks of criticism is to analyse the individuality of the artist (that is, his art) into its component elements, and to show their correlations. In this way, criticism brings the artist closer to the reader, who also has more or less of a 'unique soul,' 'artistically' unexpressed, 'unchosen' but none the less representing a union of the same elements as does the soul of a poet. So it can be seen that what serves as a bridge from soul to soul is not the unique but the common. Only through the common is the unique known'
>
> Leon Trotsky, *Literature and Revolution* (1923)

0.1 Aims

The aim of this thesis is to demonstrate the suitability of John Burrows' Delta method to the identification of changes of historical significance in literary history. It is the contention of this thesis that dialectical materialism offers an eminently suitable means of considering empirical data while retaining a rigorous conceptual framework and proposes it as a lodestar for computational literary scholars going forward. This is undertaken, not solely in order to demonstrate its usefulness, but because it offers the most robust means of charting in order to chart our way through some of the antinomies at work in computational literary studies in general as well as those structural and ideological factors which have forestalled the emergence of an authentically historical orientation within the discipline.

0.2 Research Background

The history of the application of statistics to literature can be traced as far back as the thirteenth century, to the manual compilation of biblical concordances (Cooper and Mikhailov 48). In its current form, as an academic discipline, this activity has been referred to as quantitative literary criticism, stylometry or quantitative stylistics to provide just a few examples. Computational literary studies, or CLS as this thesis will be referring to it from this point onwards, is a nascent but growing field of inquiry within literary-critical discourse and was initiated by greater amounts of access to electronic computers in research institutions in the sixties. Over the course of the following three decades, research output within CLS was almost exclusively composed of the work of a small number of scholars, working in the fields of both literature and statistics. These scholars responded, refined and engaged with one another's other's uses of mathematics in order to approach long-standing questions within literary criticism, including the derivation of empirically-based methods for identifying the order in which particular works were written (Holmes 112; Mansell), or attempts at validating established literary-critical hypotheses (Robey).

The emergence and entrenchment of what Robert Brenner, with appropriate caveats, refers to as the 'new economy' can be characterised by the growth of globalised patterns of consumption and investment (Brenner 223–26). This re-structuring of the global economy, in its movement towards transnational supply chains of production and distribution, depends to a large extent on the increasing power and capacity of communications and information technology, including, or perhaps especially, computing. These economic trends have had significant consequences for publicly funded institutions over the past few decades. Universities, for instance, are increasingly called upon to assist in the shoring up of the nation-state's international economic competitiveness, a phenomenon which Stefan Collini has identified and written on extensively (Collini 21). A symptom of these changes can be perceived in universities' increasing focus on STEM-oriented research. When the development of software, hardware or data analytical methods do not form a constitutive aspect of a particular discipline, as would be the case in the vast majority of subjects located within humanities or social science faculties, there are additional incentives attached towards the identification of research questions which may be productively engaged at least in part by computation or the analysis of big data; this phenomenon is referred to within the literature as interdisciplinary research and in its capacity to direct predominantly non-STEM subject areas into the realm of quantification, its advantages in this context are obvious (Wernli and Darbellay 5). In this sense, third-level institutions now inhabit a milieu within which they are materially invested in the generation of knowledge via the application of methods such as machine learning, artificial intelligence and other forms of applied statistics or data science.

Given these objective trends within higher education and research briefly sketched above, as well as the increasing availability of large-scale corpora of textual data available for quantitative analysis in open formats on the internet, such as Project Gutenberg (PG) and Early English Books Online (EEBO), the potential now exists for CLS to move beyond its limitations as a curio undertaken by a small number of scholars in relatively obscure academic journals and to become instead a secure fixture within the field of literary criticism. CLS may also be instrumental in allowing literary criticism to engage the broader public to a greater extent, if the coverage it seems possible for findings which arise from within CLS research projects to receive in international and mainstream media outlets represent an acceptable barometer (Radiolab; Flood).

Within any historical account of CLS, it is of course also necessary to account for the digital humanities (DH). Over the course of its history, DH has, for better or worse, been defined primarily through its lack of a secure disciplinary, institutional, or even conceptual 'centre' (Klein and Gold x; Svensson 83). Despite this lack of secure definitions, for our purposes here it will be sufficient to describe DH as the critical application of computing to research questions which have traditionally been the preserve of researchers working within humanities or the social sciences (Schreibman et al.). To say that DH has been controversial since the phrase first entered common circulation in the early to mid oughts would represent something of an understatement. To its detractors, DH represents an uncritical reflection of the political and economic trends outlined above, wherein universities and other research-oriented institutions have directed their focus more towards disciplines which offer straightforwardly commercialisable research of interest to industry and private enterprise. A number of scholars have therefore argued that DH exemplifies the culmination of a neoliberal turn in university governance; polemic disputes regarding DH's political valences are another frequent fixture of the scholarly discourse surrounding DH as a result (Greenspan). Some of the negative subtext which DH has acquired as a field over the past two decades go some way towards accounting for why an increasing number of CLS studies have attempted to re-locate their research within a more coherently literary-critical context. Rather than concerning themselves with the nature or ethics of the tools they use in order to produce knowledge, CLS scholars have begun to engage research questions of central importance to literary criticism writ large, such as the viability of locating particular works within specific taxonomies of genre as opposed to others or the appropriateness of periodising models and their relationship to historiography. We see these concerns and many others being brought to the fore in contemporary CLS scholarship such as Andrew Piper's *Enumerations* (2018), Martin Paul Eve's *Close Reading with Computers* (2019) and Ted Underwood's *Distant Horizons* (2019) and it is primarily relative to studies of this kind that this thesis locates itself.

0.3 Objectives

The central task of this thesis is to analyse a large corpus of word frequency data obtained from the HathiTrust Digital Library (HTDL). This corpus contains tens of thousands of works of prose, poetry and drama published between the years 1700 and 1922. We analyse this dataset in order to identify sudden and extensive differences in the literature produced in one period of time when compared with the output of a previous period. We refer to these posited instances of accelerated change as 'breaks.' Before providing additional background regarding the quantitative means through which a 'break' may be identified, we will consider the way in which breaks have been framed within the literary-critical discourse.

John Brenkman identifies the operation of a doctrine inherited from the avant-gardist movements of the early twentieth century within the notion of artistic innovation (Brenkman 818). According to Peter Bürger's influential account of the histories of Surrealism, Futurism and Constructivism, the *avant-garde* represent a critical response to the socially reactionary aesthetic and classical modernisms of figures such as T.S. Eliot or Henry James. The *avant-garde* therefore brings the scepticism regarding bourgeois society which we may perceive in canonical works of classical modernism to its logical end-point, rejecting the separation between art and social life *in toto* via a praxis-oriented artistic philosophy (Burger 33–34; Pinkney).

In *Marxism and Form* (1971), Jameson frames the notion that the art produced in one period of time marks a significant advancement from the art of another era in more economic terms and as emblematic of the influence of industrialisation on cultural form, as innovation comes to be valued in the same way as technical innovation is within industrial production. This is indeed an imperative, given the numbness which trends of cultural consumption under capitalism inculcate within the subject (Frederic Jameson, *Marxism and Form: Twentieth-Century Dialectical Theories of Literature* 19–21). In this sense then, the notion of cultural innovation reflects the development of the same economic and political forces seeking to grant hegemonic legitimation to the accumulation of capital on a global scale. As Jameson also notes however, we may obtain from Theodor Adorno, an alternate means of regarding this movement, in such a way which may assist in disassociating this connection. Adorno argues that what lies behind literary modernism's drive to create innovative art is not so much an uncritical cleaving to the new, but rather a sense that certain forms, expressions and techniques have become, through their overuse, overly conventional or kitsch and must be creatively avoided (Frederic Jameson, *The Modernist Papers* 5). Ástráður Eysteinsson, similarly frames the break in negative terms, locating its origins in the writings of the German philosophyer Friedrich Nietzsche which pertain to the role of forgetting in the construction of historiography. As Eysteinsson glosses this concept, the break is by its very nature a paradoxical gesture, as it necessitates a certain degree of self-consciousness regarding both what it is that is being negated and what is taken up in its stead (Eysteinsson 53–54). Regardless of where exactly this notion originates in modern history, both Jameson and Brenkman relate the orientation of the *avant-garde* to the Russian formalist notion of defamiliarisation and as integral to any understanding of the literary; a concern with the productive alienation of language which the work of art facilitates can be identified within almost all schools of twentieth century literary criticism (Brenkman 823; Frederic Jameson, *The Prison House of Language: A Critical Account of Structuralism and Russian Formalism* 45).

The degree to which any specific moment of cultural production can be said to erect a definitive and decisive movement away from all prior moments is currently out of step with the prevailing literary-critical discourse. Speaking in very broad terms to be expanded upon later in this thesis, literary history is currently defined more in terms of an assemblage of contingent and discursive relations based on perpetual exchange, within which any directional, temporal or teleological dimension is to a large extent repressed, or sublimated. This is partly the reason why, as the architectural critic Owen Hatherley notes, models of cultural history which emphasise incremental change are so fashionable in contemporary criticism and scholars tend towards the dismissal rather than the promotion of the idea of revolutionary breaks within cultural history (Doherty).

In addition to being out of step with current academic fashions, it would seem as though the break hypothesis of literary history has already been debunked within CLS. In *Distant Horizons*, Underwood convincingly demonstrates that the cohort of words which correlate, both positively and negatively with the passing of time, increase or decrease steadily over the course of three centuries. Based on Underwood's graphs and accounts of his sequence of results, there is no basis for identifying any particular point in the historical record at which they can be said to come into greater or lesser amounts of prominence (Underwood, *Distant Horizons: Digital Evidence and Literary Change* 23–25). There is, however, a disguised risk in adhering too closely to this hypothesis of incremental change. Speaking in a more polemical vein, Hatherley posits that the incremental perspective is 'such a boring way of looking at anything' (Doherty). It is in this spirit that this thesis attempts to re-introduce some notion of the revolutionary break in literary-critical history in its analysis of the HTDL word-frequency dataset from the point of view that the tradition within which Jameson and Adorno write offers the best means of doing so

Our means of identifying these breaks will be derived from Alexander T.J. Barron et al.'s temporally-based topic model of innovation and influence in the speeches and debates of the first parliament of the French revolution. Barron et al.'s method begins by calculating the distances existing between each successive speech delivered in the National Constituent Assembly (NCA) based on the degree to which each one makes use of each of the 100 topics identified via a Latent Dirichlet Allocation (LDA) topic model. A temporal dimension is then introduced to these calculated distances. Distance forwards is identified as being indicative of a speech's *transience*, distance backwards indicative of its *novelty*. These two figures are subtracted from one another in order to attain a third figure, referred to as *resonance*. On the basis of these three statistics, Barron et al. models the particular contribution each speech makes to the proceedings of the NCA. If they exist at relatively greater amounts of distance from previous speakers while being relatively proximate to subsequent speakers, their speeches are both highly novel and highly influential. We might equally expect particular speakers to have high values for novelty, but low values for resonance, indicating lower 'impact' overall (Barron et al., "Individuals, institutions, and innovation in the debates of the French Revolution."). The potential of this approach, were it to be operationalised within a literary-critical context, is clear. If successful, it would allow us understand, not only cultural change, but also its dynamics and its capacity to exert itself to a greater or lesser extent. Put simply, this method offers the potential to model literary influence itself.

As the choice of literary critics thus far into this introduction suggests, the subset of literary criticism which has most consistently approached the analysis of literature in these terms, rendering sudden changes in the literary-historical record, coherent within a *longue-durée* historical account of human history and society, is predominantly Marxist in orientation. Critics such as Frederic Jameson, Raymond Williams, Perry Anderson and Terry Eagleton have, to great effect, produced detailed analyses of specific literary works which can be regarded as symptomatic or significant within the qualitative transformations which take place within a given historical moment. With the development of large collections of works and computational methods adequate to their analysis, it is now possible to integrate empirical evidence into these Marxian models of literary history, allowing us to grasp the broader objective tendency lying behind these symptomatic readings. In making this point, we do not wish to suggest that the arguments which these critics have mounted in the past are in any sense inadequate to this task; no-one would deny the conceptual robustness, or synopticism which the critics named above have brought to their studies of cultural production without the benefit of statistics. Individual examples within literary criticism will furthermore always be to some extent necessary. As a means of rendering literary-critical hypotheses intelligible symptomatic readings are unsurpassed in their efficacy, but there is no reason why the research of the critics cited above could not be enhanced or furthered by the introduction of different forms of quantitative, or otherwise empirically-derived, evidence. This lack of empirical evidence in fact represents a significant problem for a literary-critical historiography rooted within dialectical materialism. As Vladimir Lenin notes in his account of Marx's life and works, the robustness of Marx's critique depends upon its capacity to draw from three distinct schools of inquiry. In formulating his analysis of nineteenth century political economy, Marx drew to a significant extent from empirical and statistical methods for the tabulation of data as well as the rigours of German conceptual and French utopian socialist thought

(Lenin 7); it is on this empirical and statistical front that literary-critical histories have consistently fallen short. The aim of demonstrating the possibility of a classical Marxist approach for literary-critical inquiry represents a significant component part of this thesis' intended approach.

0.4 Structure

This thesis aims to develop a point of departure for more historically oriented CLS, in following some of the methodological lines of inquiry first proposed by Burrows and developed by Ted Underwood, Jan Rybicki and Maciej Eder. It does so first by providing a quantitative and methodological history which has not yet been mapped systematically. By charting the changing contours of how CLS has developed over the past few decades, as well as by demonstrating its capacity to function in this instance, this thesis hopes to make a small but not insignificant contribution to the changing discourse within CLS and how it might yet develop within the decades to follow.

Chapter one constructs a history of CLS, with particular attention to the CLS scholarship produced by the Australian literary scholar John Burrows. Burrows' Delta method was pivotal in challenging many of the orthodox assumptions of CLS' early years, to the extent that the history of CLS may be broken into three distinct moments of roughly two decades each: i) pre-Delta (1963 - 1979), ii) the era of Delta's emergence (1980 - 1999) and iii) post-Delta (2000 - 2020). Subdividing the historical record like so not only allows for the critical evaluation of each distinct moment in and of itself, contextualised within a long-term narrative of development, but also allows us to assert the degree to which CLS is a field increasingly amenable to the consideration of literary change within an historical context. This change can be said to correlate with CLS' movement away from the consideration of individual or specific formal features in a relative historical vacuum as compared to more holistic approaches such as Delta, which in their contemporary incarnations, involve the analysis of texts more or less in their entirety. Some of the reasons why CLS took a relatively long period of time to come to this amenability to historiography, will be accounted for within the broader literary-critical history this chapter offers.

Having accounted for the current state of play within CLS in this thesis' first chapter, our second methodological chapter outlines the means through which the HTDL dataset this thesis adopts as its object of study was loaded into the R workspace, rendered suitable for the application of the Delta method and how the resulting Delta distances were modelled diachronically in order to identify breaks in modern literary production. The third chapter presents some of the results which emerged from this modelling process and how the machine learning technique, regularised logistic regression, was applied to two blocks of data existing on either side of each of the breaks which are identified, in order to identify the word types which can be can be said to both play a decisive role in transforming one literary epoch into another, as well as serving as indicators of this transformation, by making themselves felt within literary works to a significantly greater or lesser extent.

The quantitative analysis outlined in our second and third chapters largely reinforces Underwood's analysis in *Distant Horizons*, in that it conclusively demonstrates literary history advances slowly and steadily between 1700 and 1922 rather than being subject to conjunctural transformation. However, just because the incidence of a group of words increase or decrease their relative incidence over a two hundred and twenty-two year period does not mean that we cannot incorporate them into an historical account of modern literary production. Our fourth chapter therefore identifies these extracted word types and re-locates them in their proper contexts, identifying sections of literary works in which they occur to a significant extent. By evaluating these words from the perspective of the content as well as the historical periods within which they originate, we may obtain a novel insight into literary history over the eighteenth, nineteenth and early twentieth centuries. This chapter considers germane texts produced by literary authors such as Charles Baudelaire, William Wordsworth, D.H. Lawrence, Virginia Woolf and Stephen Crane. This is undertaken in the context of a broader consideration of how literary criticism's emerging methodologies, especially CLS, may be rendered coherent with influential Marxian accounts developed by Jameson, Eagleton and Williams, within the neo-liberal university at the present time. This thesis then ends with a concluding chapter which encapsulates the results obtained in the context of the study and presents some potential future directions for research in the field of CLS.

Chapter 1

Literature Review

Before we move on to consider the material itself, it is first necessary to clarify exactly what is meant by quantitative literary history. We contend that quantitative literary history requires the satisfaction of a number of criteria and is meaningfully distinct from what Constantina Stamou has referred to as 'stylochronometry' (Stamou 181), the analysis of formal features in texts or corpora over time. The development of a quantitative literary history, or an historical CLS, necessitates the satisfaction of three distinct criteria. Firstly, reliable metadata relating to date of publication or composition. Secondly, a broad digital corpus of literature, in the tens or hundreds of thousands spread over at least a century. Examples of corpora of this size may be Early English Books Online (EBBO), or the HathiTrust repository, which this thesis takes as its object of study. The point is that these collections capture an approximate or reasonable proportion of the publishing output of a particular period and that an analysis into their contents could be said to constitute a computational and literary study of historical significance. Thirdly, the development and calibration of quantitative methods capable of analysing texts in their entirety, rather than identifying or enumerating specific formal features, such as a cohort of words, combinations of words, or punctuation marks alone, in order to satisfy a particular literary-critical hypothesis.

This chapter will seek to argue that, for a number of reasons, the analysis of individual or specific formal features, as in univariate analyses, can be said to characterise CLS when it first emerged as a field of inquiry in the sixties and seventies. The tendency to focus on isolated formal features in the discipline's early days, is symptomatic of CLS' inclination to reverse the death of the author at the hands of figures such as Roland Barthes and Michel Foucault, re-emphasising the individual agency and style of the author. In its early history therefore, CLS retains more romantic theories of authorship, a tendency which results in the development of methods which assume that all texts written by different authors are differentiable on the basis of parameters which are in fact wholly arbitrary. This chapter will trace the movement of CLS away from univariate methods to multivariate methods more adequate to broad, historical analysis for reasons of methodological as well as conceptual accuracy. It is in the eighties and nineties that we first see the previously regnant univariate methods consistently outperformed by multivariate approaches in which approximately 100 of the most frequent words (MFWs) in a text are quantified. In the oughts and tens CLS scholars extend these methods further; the success of John Burrows' Delta method lays the foundation for the analysis of thousands of words and the treatment of texts more or less in their entirety. This therefore marks the point in time in which romantic ideas of authorship within CLS began to be challenged and the possibility of an historical CLS, particularly one which can draw from developments underway in the context of machine learning, begins to emerge.

In accounting for these three phases in the history of the development of historical CLS, this chapter will be divided into three parts, emphasising particular works of scholarship which have been instrumental in transforming one epoch into the next. The first two of these sections, 'Embryonic CLS' (1963 - 1979) 'PCA & Proto-Delta' (1980 - 1999) will be in turn divided into three sections: i) multivariate methods, ii) historical CLS and iii) literary theory. While this chapter's third and final section retains a section on literary theory, the other two are jettisoned in favour of a section on i) Delta and ii) machine learning. This difference in structure is due to the changing character of CLS research produced after the year 2000. From this point onwards, the amount of scholarship produced in the context of CLS increases significantly. The history of the discipline is therefore a somewhat lopsided one and the rate of change within CLS can therefore be said to have accelerated significantly from this point onwards. One of the consequences of this is that the boundaries between these formerly relatively discrete strains have begun to become blurred. Questions of computational literary historiography begin to be considered to a greater extent within more theoretical sections of the literature and machine learning, only occasionally present in CLS research undertaken before the year 2000, begins to become customary. In short, the use of multivariate approaches had become more or less a *sine qua non* of all CLS research and changes which come over this chapter's structure aim to reflect that.

Any chronology which accounts for the discipline's history will be a generalising one and will require the omission or simplification of particular phenomena. Some articles anticipate transformations within the discipline which are later to take place and, as we will also see, CLS scholars are sometimes prone to continuing to use methods which have been shown to be inadequate elsewhere. The passing of more time and the creation of more research will be necessary before the full significance of these changes become clear. Nevertheless, the periodisation here proposed allows us to introduce both superstructural and infrastructural causes in considering the discipline's history.

1.1 Embryonic CLS (1963 - 1979)

1.1.1 Multivariate Methods

As Jack Grieve notes, there is a long history of mathematics being brought to bear on the study of attributing authorship, reaching back to the nineteenth century (Grieve 251). However, in his history of the field, David Holmes identifies the first instance of modern stylometry in Frederick Mosteller and David Wallace's attempts to identify authorship in the twelve pseudonymously written essays and articles in *The Federalist Papers* (1788) written by Alexander Hamilton, James Madison and John Jay. Mosteller and Wallace attribute authorship on the basis of similar rates at which function words are deployed in the text, such as prepositions, conjunctions and articles (Holmes 112). Fred Damereau seems to be the first to account for the use of function words from a theoretical perspective, citing W.J. Paisley's theory of 'minor encoding habits.' According to Paisley, in turn drawing from theories developed in the field of art history, indices of personal style can be found in minor, but highly common, features of a work. They should not vary significantly between works produced by the same author but should vary significantly between works produced by different authors. In satisfying these criteria, Damereau identifies function words as being most suitable (Damerau 271–72). Damereau's approach is perfectly logical given that it is wholly appropriate to reduce most authorship attribution problems to what Burrows refers to as a 'closed game.' When there is a restricted set of texts and candidate authors, function words seem to be capable of providing promising results. However, this fact culminates in the assumption that authorship is in and of itself a guarantor of a distinctive or individual style which suffuses the work in its entirety. It is therefore assumed that each text produced by a single author is statistically homogenous and that any given quantity of features identified in a text written by one author will be statistically distinct from the same feature in a text written by another author, under the assumption that this can be confirmed through the use of Mann-Whitney, chi-square, Student's *t* and Fisher tests. As this chapter proceeds we will see that this assumption has been influential within the context of many other CLS studies, that its influence is detrimental and compounded by, as David Holmes notes, the lack of a methodological 'holy grail' within stylometry which can operate in all contexts, genres, languages or eras (Holmes 111). Barron Brainerd's work, in its capacity to identify and willingness to test the resilient assumption of intra-authorial heterogeneity, represents an exception and Brainerd is therefore among the first to identify some of the drawbacks associated with the use of the chi-square method when applied to literary texts (Barron Brainerd, "Statistical analysis of lexical data using chi-squared and related distributions" 161; Barron Brainerd, "Pronouns and genre in Shakespeare's drama" 5–12).

There are a significant number of papers in *Computers and the Humanities*' (*C&H*)'s early history which abide by sound statistical and methodological practice, such as Paule Sainte-Marie et al.'s application of principal component analysis (PCA) to 44 MFW's in 30 plays written by Molière (Sainte-Marie et al. 136) and Brainerd's application of cluster analysis in order to differentiate novels from romances (Barron Brainerd, "On the distinction between a novel and a romance: A Discriminant Analysis" 267). However, many other *C&H* and *Literary and Linguistic Computing* (*LLC*) articles until the nineties can be characterised by the arbitrariness of their methodological approaches. Sampling, variable selection and statistical measurements are often adopted and applied without explicit reasoning or reference to previously undertaken studies within which the efficacy of these methods have been validated. Citations are also less common in early articles than they later become, and this has the effect that the precise rationale for any given procedure being carried out is more often assumed than explained. Robert Cluett's analyses of part-of-speech (POS) entities in Restoration-era prose and John Foley's analyses of stress patterns in Beowulf, represent another tendency rife at this early stage in CLS history, which take a heuristic approach to drawing conclusions rather than using proven mathematical techniques (Cluett 264–68; Foley 78). These defects can probably be accounted for by bearing in mind the nascency of the field. As M.W.A. Smith notes, at time of writing in 1987, there was no extant corpus of studies undertaken which had successfully inculcated an understanding of statistical best practice when analysing literary texts and CLS scholars could not benefit from a corpus of articles on which to base their approaches in the same way a would-be CLS scholar could today (Smith, "Hapax Legomena in Prescribed Positions: An Investigation of Recent Proposals to Resolve Problems of Authorship" 145–46). Other constraints which exert a significant influence on the early scholarship include the available infrastructure. The computing capacity of any given machine, which would have been a factor in early experimental design, go some way also in explaining the methodological focus we see on quantifying the frequencies of a very small number of function words. Computing hardware and its use was also expensive and prior to the sharing of digital texts via the internet each researcher would need to build their own corpus (Sainte-Marie et al. 131–32; Sula and Hill 191).

1.1.2 Historical CLS

It is in the context of analyses of manuscript data that we begin to see the first inklings of an historical or contextualist impulse beginning to emerge, as in Darrell Mansell's attempts to date the composition of different sections of Ernest Hemingway's novella *The Old Man and the Sea* (1952) based on differences in the number of syllables per word (Mansell), a method which seems to have been derived from D.R. Cox and L. Bradwood's attempts to construct a chronology of Plato's works based on quantitative evidence (Holmes 112). Finally, there is a tendency within CLS which is best encompassed by a consideration of J.M. Coetzee's study of Samuel Beckett's *Lessness* (1970). Coetzee identifies the second half of Beckett's work as a series of iterations through segments of the first, thereby demonstrating that the work formulates itself more as a word game than verbal expression as such, concurrent with post-structuralist theories of literature with which Coetzee would have been engaging at the time (Coetzee). Robert M. Henkels, Jr. and Esteban R. Egea's attempts to identify syntactical repetition within Robert Pinget's *Fable* (1970) which would correspond to the structural repetition which Henkels et al. identify in Pinget's writings (Henkels Jr and Egea) represents another manifestation of this same impulse. The commonality between these two articles is that they both represent attempts to operationalise literary-critical theories in a computational setting, despite the fact that validating these hypotheses is arguably beyond the capacity of a suite of statistical methods to validate. These analyses are more invested in using quantitative methods as a means of mediating the details of a work with a more macro perspective. Though these analyses often produce interesting and productive results, they tend to take their lead from what is currently possible or predominates within the field, rarely than advancing it in any significant way. Nevertheless, they represent an important tendency, which we see consistently throughout its history and it is therefore necessary that they be considered as this chapter continues.

1.1.3 Literary Theory

The early polemics and considerations for the prospects for a future CLS which we find in the first issues of *C&H* are illustrative as regards the 'theory wars,' a consistent fixture of CLS discourse. It is Louis Tonko Milic who initiates this dialogue, both in *A Quantitative Approach to the Style of Jonathan Swift* (1967) and in two articles which argue for the significance and contributions computing may potentially make to the study of literature. Milic's arguments are based on the capacity of computing to alert the critic or analyst to patterns and trends which are not detectable via traditional, qualitative approaches. This is particularly important from Milic's perspective, as words which are traditionally deployed in the interrogation or analysis of style in literary criticism are vague or impressionistic. Milic partly attributes this to the blurring of the boundary between literary criticism and social theory (Louis Tonko Milic 27–28, 38, 54). In solving this problem, Milic wished to facilitate a synthesis between computation and the creative intuition which has historically predominated within literary criticism rather than automate the latter out of existence (Louis T. Milic 5). Milic begins from the notion that syntax may provide a deep and unifying structure or promising starting point for quantitative approaches (Louis Tonko Milic 32, 79) and proceeds by dividing words into twenty-four different grammar-types, looking at how the means of these word-types increase or decrease in Swift's writings over time. Milic then carries out close readings of these grammar types in their context within the works (Louis Tonko Milic 174, 205, 272).

It is Emmanuel Mesthene who presents the first sceptical response, arguing that for all the precision and accuracy which computational tools have the potential to introduce, they will introduce bias to literary-critical research as computing cannot serve as a neutrally clarifying agent (Mesthene 2). Bruce A. Beatie cites C.P. Snow's essay 'The Two Cultures'

(1959), in order to locate literary studies within a school of thought totally opposed to that of statistics (Beatie) while Susan Wittig objects to CLS on the basis of a more overt commitment to post-structuralism, which envisions the text as an ineffable system of exchange which resists all forms of hierarchical categorisation (Anderson, *In the Tracks of Historical Materialism: The Wellek Library Lectures*). This is utterly contrary to the ways in which natural language processing (NLP) and linguistic analysis requires us to regard text (Wittig).

Despite being written more than half a century ago, these four critics broadly anticipate the two opposed positions we now confront in considering CLS' relationship with the broader literary-critical milieu, even to the present day. Milic, on the one hand, emphasises the capacity of computation to allow the critic to exceed their individual point of view and potentially gain access to an hypothesised deep structure, while CLS' detractors mount an overall objection to CLS in principle, refraining from engaging with statistical methods themselves or a history of their application on the basis that empiricism is an inveterately instrumentalised and insufficiently reflexive form of knowledge production. As this chapter continues, we will see that these two positions and the tensions residing within them are crucial to any account of CLS history.

1.2 PCA & Proto-Delta (1980 - 1999)

1.2.1 Multivariate Methods

In the eighties we see John Burrows publish analyses that anticipate the Delta method he would later develop and which this thesis adopts as central to its approach. Burrows begins by focusing on the changing rates at which modal auxiliaries are used in six novels written by Jane Austen over the course of her life (J. F. Burrows, "Modal Verbs and Moral Principles: An Aspect of Jane Austen's Style"). Though Burrows argues his approach allows for the treatment of texts in their entirety, against literary criticism's historical tendency to focus on highly specific features of a work, such as modal auxiliaries and how they relate to sentence length, Burrows remains constrained within the framework he aims to supersede (J. F. Burrows, “Modal Verbs and Moral Principles: An Aspect of Jane Austen’s Style” 20–23). In his second article in *C&H*, Burrows attempts to quantitatively differentiate three different narrative categories which he identifies as being at work in Austen’s novels; i) dialogue, ii) ‘pure narrative’ — here meaning the voice of the narrator alone — and iii) ‘character narrative,’ here meaning the voice of the narrator mediated by the thoughts or feelings of a particular character, elsewhere referred to within literary criticism as free indirect discourse. Burrows first correlates the frequencies of a list of function words which appear in each of these three categories, then applies a statistical transformation to these correlation coefficients, which is referred to as eigenvector analysis or PCA. Burrows applies this method in a series of distinct permutations, firstly separating the three different narrative types by gender, then by character, describing each time the clustering patterns which can be observed relative to the literary-critical discourse surrounding Austen (J. F. Burrows, “Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style” 64–69). In his third article, Burrows applies his method to fifteen other nineteenth century novelists. As before, Burrows is invested in identifying a unique and individual style for each author and though his graph has no temporal component, he argues that each author’s oeuvre clusters chronologically and that Austen, George Eliot and Elizabeth Gaskell’s relative distance from the other authors justifies reading their styles as individual, moving away from neo-classical prose styles which otherwise predominated in the late eighteenth and early nineteenth centuries (J. F. Burrows, “‘An ocean where each kind. . .’: Statistical Analysis and Some Major Determinants of Literary Style” 318; Holmes 113). Burrows’ judgement that Austen, Eliot and Gaskell all possess distinct writing styles is largely borne out in the results of a PCA applied to the 50 MFWs which can be seen in *Fig.* 1.1.

The words which appear on in *Fig.* 1.1 on the end of arrows indicate the decisive variables in terms of determining the ways in which the data points scatter. It seems as though differences in the use of the word ‘and’ is decisive in separating the prose of Gaskell and

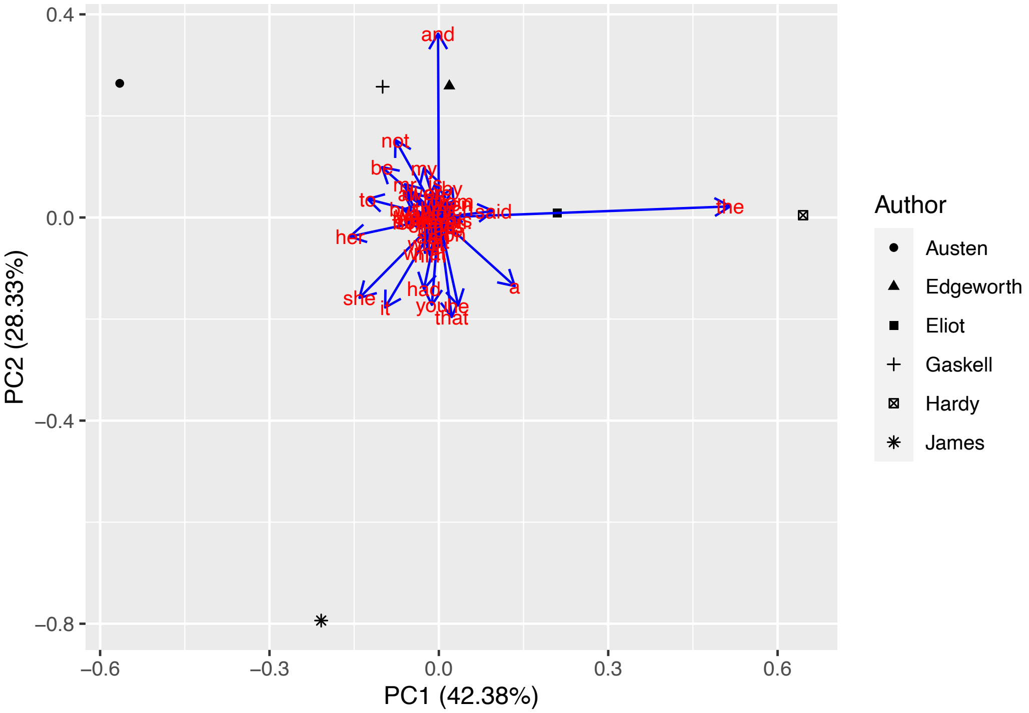


Figure 1.1: Principal Component Analysis (50 MFWs)

Edgeworth, that uses of the word 'the' is a key index of Thomas Hardy's style as well as to a lesser extent George Eliot. It will be noted that the first component, PC1, which accounts for 42.38% of the variation, is graphed along the *x*-axis and the second, PC2 accounts for 28.33% of the variation and is graphed along *y*-axis. Combined, these two components account for just over 70% of the variation overall.

Given their demonstrated capacity to cluster texts on the basis of authorship, genre and era, function words remain central within CLS and we see a number of studies emerge which continue to demonstrate the efficacy of the method (J. F. Burrows, "Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information"; D. H. Craig; H. Craig; Tse et al.). We also see further interrogations of PCA in and of itself in Binongo et al. and Penelope Gurney et al.'s investigations into its mathematical principles (Binongo and Smith) and limitations. Penelope J. Gurney and Lyman W. Gurney demonstrate that PCA applied to MFWs significantly outperform attempts to attribute authorship on the basis of vocabulary richness, a statistic which is calculated by dividing the number of unique word types by the number of words in the text overall (Gurney and Gurney, "Authorship Attribution of the Scriptores Historiae Augustae"). This result is replicated by Fiona Tweedie et al., who note that even measurements for vocabulary richness which are independent of text length are unsuccessful in discriminating texts on the basis of their authorship (Tweedie and Baayen). This is due to the fact that the richness of any given text's vocabulary is highly correlated with text length, for the logical reason that a shorter text will have far more unique word-types than a longer one. Attempts to identify a length-independent means of quantifying a text's lexical richness is a consistent fixture of CLS discourse, as in Philippe Thoiron's diversity or entropy-based method (Thoiron) or John Baker's attempts to quantify the pace at which new vocabulary enters a writer's work (Baker). The centrality of vocabulary richness to CLS may be attributed to theories of intra-authorial heterogeneity, but also the measure's relative simplicity and comprehensibility. This is probably also the case for the persistence of measurements based on sentence, word and syllable lengths, which are also plagued by similar issues relating to reproducibility (Aoyama and Constable). Gurney and Gurney recommend incorporating more MFWs into future analyses, computing space allowing (Gurney and Gurney, "Authorship Attribution of the Scriptores Historiae Augustae"). In line with the growing popularity of PCA, Robert Forsyth et al. apply PCA to the variables letter, word, doublet and string frequency, finding that all five of these parameters return an accuracy of ~70% (Forsyth and Holmes) while Douglas Biber performs a series of analyses on large corpora in order to investigate how long texts need to be represented in order to be suitable for quantitative analysis, that is, what the salient parameters are and how many texts are sufficient to characterise a genre quantitatively (Biber). Harald Baayan et al. demonstrates that the incorporation of syntactical mark-up improves the rates at which accurate attributions for authorship can be obtained (Baayen et al.), while Forsyth et al. investigates the extent to which PCA can allow for the clustering of texts written in inflected languages (Forsyth et al.). Gurney and Gurney demonstrate some of the potential pitfalls associated with clustering texts which have been divided into samples as opposed to the entire population, as different parts of texts can cluster very differently (Gurney and Gurney, "Subsets and Homogeneity: Authorship Attribution in the Scriptories Historiae Augustae").

Concurrent with the development of reliable multivariate statistical techniques in CLS, we also see previously regnant methods challenged for their failures to operate reliably. Thomas Merriam, for example, demonstrates the unreliability of 'proportionate pairs,' a method used by A.Q. Morton, which assumes that particular pairs of words which exist in a fixed ratio to one another between texts are suggestive of shared authorship. Merriam demonstrates that more than random variation can often be observed in works produced by the same author (Merriam) while Michael Hilton and David Holmes demonstrate the inadequacy of another method developed by Morton, wherein the incidence of two formal features are plotted on a line graph. The two lines are then superimposed on one another and it is determined that any instances in which these lines deviate from one another are indicative of the intervention of a second author. Hilton and Holmes propose a more statistically rigorous variant of this approach, which incorporates the weighting of particular features, but concludes that even with these improvements, they fail to reliably attribute authorship (Hilton and Holmes; Holmes 114). Smith also publishes a number of articles which challenge the use of chi-square tests, on the basis that they are prone to delivering false negatives (Smith, "An investigation of Morton's method to distinguish Elizabethan playwrights") as well as Morton's correspondence analyses, based on obtaining corresponding values of particular words in particular positions and collocation analyses, which quantify occurrences of a prescribed word either followed or preceded by a second prescribed word (Holmes 113; Smith, "A Critical Review of Word-links as a Method for Investigating Shakespearean Chronology and Authorship" 202–03; Smith, "Hapax Legomena in Prescribed Positions: An Investigation of Recent Proposals to Resolve Problems of Authorship"). Smith goes on to criticise CLS scholars for using methods which are insufficiently rigorous and proposes instead analysing the rates at which the first word in every speech appears per 1000 words in the works of six Elizabethan-era playwrights. Smith demonstrates his method's capacity to correctly identify John Webster as the most likely candidate of the six to have authored *The Duchess of Malfi* (1614) and Ben Jonson as the most likely to have authored *The Alchemist* (1610). On the basis of the seeming capacity of Smith's method to function, Smith proposes George Wilkins as being the most likely to have authored *Pericles* (1619) (Smith, "The authorship of Acts I and II of Pericles: a new approach using first words of speeches"). Later on in the period 1980-1999, we see studies which continue to draw from discredited approaches such as the chi-square tests (McColly) and the inspection of visualisations (Anderson and McMaster, "Quantification of rewriting by the Brothers Grimm: A comparison of successive versions of three tales"; Irizarry, "The two authors of Columbus'Diary") but these increasingly represent the exception. Even in instances in which PCA is not deployed, in favour of more generic visualisation of distances, analyses employ increasing numbers of variables (Greenwood, "St Paul Revisited–a Computational Result"; Greenwood, "St Paul Revisited–Word Clusters in Multidimensional Space";

Irizarry, “One Writer, Two Authors: Resolving the Polemic of Latin America’s First Published Novel”). In their attempts to attribute a number of articles and essays to Oliver Goldsmith for example, Peter Dixon and David Mannion calculate the distance between the texts Goldsmith is known to have authored and each of the disputed texts on the basis of sentence length, length of the final word in a sentence, proportion of sentences ending in a verb, to give only a few examples. Dixon and Mannion then come to a decision regarding what the ‘critical distance,’ the point beyond which a text could not plausibly be attributed to Goldsmith, might be (Dixon and Mannion). Another example can be seen in both M Stratil and R.J. Oakley’s as well as Stephen Usher and Dietmar Najock’s attempts to identify the authorship of disputed texts, clustering them on the basis of a number of high-frequency word types and a number of different distance measurements (Stratil and Oakley; Usher and Najock). While approaches such as these do not yet consider the advantages of using thousands of MFWs as is common today, they still represent the gradual movement of CLS towards holistic analyses of text and a heterogenous number of quantitative methods and away from impressionistic methods developed under the assumption that all authors possess a unique writing style which suffuses all features they deploy within a text to the extent that every text produced by any given author can be differentiated from any text produced by a different author.

1.2.2 Historical CLS

The historical or contextualist aspects of CLS at this time are overwhelmingly concentrated on dating the composition of Shakespeare’s plays (B. Brainerd) or attempts to chart an author’s changing styles over time. These tend to approach the question either according to high (Craik and Kaferly; Temple) or medium-frequency features (Laffal). However Roseanne Potter demonstrates a contrasting interest in historical and contextual approaches in and of themselves in her application of analyses of variance (ANOVA) to POS entities in English, Irish and American plays (Potter, “Toward a syntactic differentiation of period style in modern drama: Significant between-play variability in 21 english-language plays”). Content-based CLS approaches continue, as in Stephane Hogue et al.’s use of correlations in order to access a particular word’s ‘semantic field’ in the writings of the Danish philosopher Søren Kierkegaard (Hogue and McKinnon) and Paul Fortier’s investigation into the themes of French novels (Paul A. Fortier).

At this time, content-based CLS approaches seem to lend themselves to the investigation of structuralist readings as in Gregory Lessard et al.’s analyses (Lessard and Hamm; Lessard and Bénard) or Ira Nadel et al.’s diagrams of grammatical flow in Shakespeare’s sonnets (Nadel and Matsuba). The formal features which are quantified in operationalising these readings vary; we see the correlation of POS tags, sentence lengths, stresses, rhythm and other morphological features with the names of particular characters or thematically significant words (Logan and Logan; Oostdijk; Roberts; Opas and Tweedie). Some studies combine objective and subjective measurements, based on human judgement data. Some examples include Hideki Kozima’s investigation into lexical cohesion, how sections of a text may be divided into coherent segments, or C.W. Anderson et al.’s investigation into defamiliarisation in nineteenth and twentieth century poetry (Anderson and McMaster, “The Emotional Tone of Foreground Lines of Poetry in Relation to Background Lines”; Kozima and Furugori). Colin Martindale et al.’s attempts to investigate the capacity of content words to attribute authorship, an attempt which is unsuccessful in matching or exceeding the success of function words (Martindale and McKenzie) or Robert Hogenraad et al.’s investigations into the dangers of autocorrelation in content analysis (Hogenraad et al.), represent instances in which CLS is advanced in a manner more methodological than literary.

It is in the nineties that we begin to see increasing numbers of studies consider the role which machine learning might play within CLS, as in James Benson et al.’s application of a Bayesian classifier to G.K. Chesterton’s parodies of Algernon Charles Swinburne and William Butler Yeats, in order to identify whether or not Chesterton was successful in imitating their use of content words (Benson and Brainerd). Forsyth’s use of machine learning achieves a 90% success rate in assigning 142 Yeats poems to either ‘early’ or 'late' Yeats on the basis of the frequencies of random strings, representing an early instance in which machine learning methods are used to answer research questions which introduce a temporal or contextual dimension (Forsyth). Bradley Kjell and Tweedie et al. enact a number of approaches to the *Federalist Papers* via character gram and function word frequency (Kjell; Tweedie et al.). Robert Matthew and Thomas Merriam utilise a neural network, a machine learning technique which simulates biological brain function by modelling the dynamic relationship existing between a set of input and output signals (Lantz 236–40), in order to differentiate between the works of Shakespeare, John Fletcher and Christopher Marlowe. Matthew and Merriam use their results to draw attention to the tough cases, which might be said to represent either instances of shared authorship, or instances in which the work of one of these author's changes to an extent which is more or less approximate to the other author (Matthews and Merriam; Merriam and Matthews). These machine learning methods tend to be deployed in a way which obscures their actual functionality; CLS scholars do not expend significant amount of time examining the actual functionality of the algorithms. The emphasis is more often placed on the algorithm's capacities to identify an optimal number of classes already identified at the outset and, as we have already seen, locating the origin of these classes in a small number of parameters or variables. In this sense, machine learning methods are used in more or less the same way as PCA is, as a dimension reduction method by another name, rather than grappling with the capacity of the method in and of itself, further examples of which we will see in the next section.

1.2.3 Literary Theory

Criticism of CLS in this era continue to maintain the inadequacy of scientific methods operationalised within literary criticism. Both Roseanne Potter and W. van Peer argue that literary studies weigh evidence in a way which is qualitatively distinct from statistics, which by necessity, requires overlooking the process-like nature of literary expression (Potter, "Literary criticism and literary computing: The difficulties of a synthesis" 94;

Peer 303). The difficulty in providing an account of these debates is that, neither side, whether they happen to be invested in maintaining a strong post-structuralist current within literary criticism or CLS scholars who wish to render literary studies more empirical, are interested in clarifying or examining what the other side is doing. This takes place to the extent that it is difficult to identify the legitimate failings of one school to which the second may provide some form of redress, let alone a new synthesis of the two positions. We need only consider Fortier's arguments that post-structuralist approaches to literature have moved beyond 'sense and reason' (P. A. Fortier) or Milic's that postmodernism, as manifested within the strain Milic regards as responsible for the death of the author, is nothing more than a mixture of 'victimisation theory,' and 'Marxism' (L. Milic 394) to identify how much more heat than light has been generated in CLS scholars' engagements with literary theory. Though the milieu at this time would seem to be ripe for the contribution of a scholar versed in both the history of statistical methods and continental philosophy as deployed within literary criticism from a synoptic point of view, in such a way which could propose a synthesis within which both intellectual endeavours would be mutually enhanced, such a work unfortunately never materialised. Rather, the straw-man argument which roughly equates one to reactionary politics and the other to an incoherent admixture of feminism and relativism, remains rife.

The closest this comes to taking place arises from within a particular school of literary criticism which runs parallel to CLS and coincides with the rise of web 2.0 technologies. A number of literary critics and authors such as Shelley Jackson, Michael Joyce and George P. Landow, begin to engage in debates surrounding textualist theories operationalised within an information technology rubric arguing that the fusion of textualist theories of literature and information technology have the potential to revolutionise literary-critical pedagogy and scholarship. Key to Landow's theories are the intersections between text containing physical links and avant-garde literature. Kathryn Sutherland for example argues that Charles Dickens' novels are hypertextual as they introduce a sense of the aleatory and contingent (Landow 185–86; Sutherland). It is important, in this context, to draw distinctions within poststructuralism, given how rare they otherwise are in CLS scholarship. While Landow's formulation of the reading experience as being akin to the subject traversing the text, choosing to engage or not to engage with maps, images and academic glosses in an indefinitely postponed arrival at closure would seem to recall the play of signifiers as we might encounter it within the work of Jacques Derrida, the scholarly and theoretical literature which develops around hypertext is more accurately rendered as a utopian reading of Giles Deleuze's and Félix Guattari's concept of 'desiring-production.' Desiring-production is a composite of Freudian psychoanalysis and Hegelian Marxism, which regards the desire of the subject as a productive and positive relation, rather than the negative valence it takes on in other branches of idealist or poststructuralist thought (Deleuze and Guattari 12). It should be noted that while Milic's intervention, mentioned above, is unsurpassed in its virulence as far as either side is concerned, there is no shortage of reductionism at play in anti-statistical arguments, as we see in Mark Olsen's argument that humanities computing is inveterately conservative in its introduction of a naive empiricism to the field (Olsen). Nancy Laan is on firmer ground when she attempts to challenge the assumption within CLS that each author employs their own style unconsciously and that this style remains consistent throughout the author's life. Laan argues that assumptions such as these are contradicted by other studies which demonstrate that an author's style does change as their career goes on (Laan). The more clearly expressed contradiction within CLS which Laan here only circles, is that these problems are symptomatic of the lack of an objective and universally applicable benchmark in CLS, which is, given the iterative nature of both literary criticism and statistical methods, unlikely to arise anytime soon. While it may be true that, for example, Austen's use of modal auxiliaries changes as her career goes on, this difference in modal auxiliaries is not sufficient to render Austen's novels indistinguishable from, for example, the novels of Elizabeth Gaskell; germane variables disappear or re-appear depending on the nature of the research question being engaged and the corpus analysed. Despite the significant advances which are made evident in the application of distance measurements or dimension-reduction techniques to increasing numbers of features, CLS at this time still seems unable to find a ground on which the literary and empirical aspects of their approach may be combined.

1.3 Delta, Results and Prospects (2000 - 2020)

1.3.1 Delta

Burrows first presents the Delta method in 2001 in an attempt to move CLS beyond the quantification of authorship from within the context of the closed game, wherein only two or three authors may be presented as probable candidates within an analysis. The Delta method's capacity to incorporate large numbers of authors, Burrows contends, will allow for the development of CLS analyses which do not close off potential avenues of interpretation before the analysis has begun. Burrows' first use of the Delta method begins by identifying 30 MFWs, disambiguating some of his chosen MFWs on the basis of their grammatical function and expressing each MFW's frequency as a percentage of the number of words in the text overall. The distribution of each word is then normalised as a *z*-score, such that each frequency is expressed in terms of the number of standard deviations it resides from the mean. The 'Delta score' is the mean differences which exist between each word's normalised frequency. A fully-worked example of how the word frequencies in a text are relativised, normalised and the distances between them calculated, albeit within the context of a subsequent modification rather than Burrows' initial formulations, can be seen in the first appendix to this thesis.

Through use of this method, Burrows demonstrates that works by John Milton are less dissimilar to one another than they are to the works of twenty-four other seventeenth-century English poets. Burrows tests Delta with 150, 120, 100, 80, 60 and finally 40 MFWs, observing a decrease in attributional accuracy with each decline in quantified MFWs (J. Burrows, "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship"). In an article published in Blackwell's *Companion to Digital Humanities*

Burrows analyses forty seventeenth and eighteenth century poems, dividing his 150 chosen MFWs into three groups based on subjective readings of their function and applies Delta to each of them separately, trying to identify which of the three cohorts could be considered to be more denotative of authorship as compared with genre (John Burrows, "Textual Analysis"). Burrows' continuing attempts to hone in on smaller-scale qualitative readings provide further evidence of how the central insight of Delta as a CLS method is continually overlooked.

Before use of the Delta method was taken up to a significant extent, Hoover published a number of articles which involve the application of distance measurements to word frequencies, albeit without normalising or relativising them. Hoover compares how rates of successful attribution are changed by altering the number of MFWs, sample size, methods of computing distance, or removing dialogue, pronouns or texts with a first person narrator, from the analysis. Hoover's analyses replicate Burrows' most significant overall finding, that the quantification of more MFWs increases the rate at which a text is successfully attributed and the most frequent bigrams such as 'it is,' 'to the' and 'of the' may be even more effective in this regard (Hoover, "Statistical Stylistics and Authorship Attribution: an Empirical Investigation"; Hoover, "Frequent Word Sequences and Statistical Stylistics"; Hoover, "Frequent Collocations and Authorial Style"). Burrows' and Hoover's analyses also offer at least one other significant insight. As G. Bruce Schaalje et al. demonstrate, Delta does not quite allow CLS to definitely break from the problem of the closed game. By virtue of the way in which Delta operates, it in fact tends towards the generation of false positives if it is applied as as a means of attributing authorship (Schaalje et al.). Scholars such as Patrick Juola have suggested a means by which Delta's tendency to do so can be reduced, by introducing a distractor corpus of true negatives, thereby raising the bar of similarity required if a text is to be identified as the most similar to any other (Juola). Even if Juola's proposed adjustment is successful, the central problematic remains in place and is in fact implicit in Burrows' initial terms of reference. Delta is thus best conceived as a means of analysing style in relational terms, than as a means of settling instances of contentious authorship. Yet it is a peculiarity of the discourse that Delta's capacity to consider style in this manner is not considered to any significant extent. We see this again in the context of two studies undertaken by Hoover. Hoover is firstly reticent to incorporate additional words into an analysis at all, on the basis that this will lead to the quantification of formal features which are within the conscious control of the author (Hoover, "Multivariate Analysis and the Study of Style Variation"). In a second study, Hoover attempts to improve attributional success by removing textual features, such as contractions or personal pronouns from the analysis and then applies Delta to one or two texts divided into a number of different parts in order to see if Delta will cluster them with one another. By his reticence to incorporate additional words into his analysis Hoover's methods therefore again attempt to return to smaller-scale qualitative readings which emphasise the decisive impact of specific formal features (Hoover, "Testing Burrows's Delta"). However, the ongoing influence of Paisley's theory of minor encoding habits demonstrates why the best means of accounting for why it is that the results of Delta analyses are so consistently passed over, despite the efforts of scholars such as Mannion and Dixon which dispute Hoover and others' focusing on unconscious formal features in favour of understanding some other features as being consciously deployed (Mannion and Dixon).

Karine van Dalen-Oskam and R.J. Stapel find that Delta is capable of detecting stylistic shifts when it is applied to medieval manuscripts (Stapel; Dalen-Oskam). Hoover et al. attempts to attribute authorship when we only have concrete evidence of one of the candidate's authorship or other non-ideal circumstances where data is lacking (Hoover and Hess; Hoover, "Simulations and difficult problems"). Hoover is the first analyst who aims to optimise Delta by making quantitative adjustments to Burrows' original method. Hoover does so by treating positive and negative z-transformed relative frequencies differently, either by focusing on higher values, or squaring and summing positive and negative means in a number of different permutations. None of these approaches are successful in outperforming Delta outright (Hoover, "Delta Prime?" 477–95) but these proposed modifications are still widely applied and compared with one another (Holmes and Crofts). Daumantas Stanikunas et al. apply a further modification known as Eder's Delta, which applies weights to frequencies in order to moderate the influence of infrequent word-types (Stanikunas et al.). Shlomo Argamon also attempts to improve Delta on mathematical grounds. Argamon points out that Burrows normalises word distributions by the mean and standard deviation, an approach which would only make sense if the word frequencies were distributed normally, but applies a Manhattan distance, which assumes a Laplace distribution. Stefan Evert et al., in a subsequent publication which systematically assesses Delta's performance with that of its subsequent improvements, confirms that based on results obtained from both English and German reference corpora, word frequency distributions are better represented by a normal than by a Laplace distribution. Given this statistical oversight, Argamon proposes three improvements. The first is Linear Delta, which retains the Manhattan distance but normalises the relative frequencies according to the median and spread. The second is Quadratic Delta, which retains Burrows' method of normalising, but applies the more mathematically sound Euclidean distance to the word frequencies. The third adjustment, Rotated Delta, is proposed on the basis of Delta's dubious assumption that word frequencies are independent. It performs a whitening transformation on the word frequencies in order to render them independent from one another (Argamon). Despite their greater degree of mathematical legitimacy however, Argamon's approaches do not outperform classic Delta (Evert et al. 8; Jannidis et al.).

Peter Smith et al. argue that on the basis of the assumptions which Euclidean distance makes, and the fact that its accuracy decreases as dimensionality — i.e., the number of MFWs we apply the distance measurement to — increases (Smith and Aldridge) that there may therefore be an upper limit beyond which we should not quantify words when conducting a Delta analysis. Smith et al. propose 200 - 300 MFWs as this upper limit, though as Fotis Jannidis et al. argue, this figure is probably quite low, and a product of the fact that Smith et al.'s study was based on an analysis of a corpus of poetic texts (Jannidis et al.). Jacques Savoy's study, which applies Kullback-Leibler divergence, Burrows' Classic Delta and chi-square in a bid to identify the optimal number of differentiators, argues for between 300 to 500 terms (Savoy). Evert et al., in demonstrating that cosine distance outperforms classic Delta, notes that Classic Delta's efficacy peaks at ~1000 - 1500 MFWs and thereafter manifests more erratic behaviour, whereas the version based on cosine distance plateaus (Evert et al. 14). Jan Rybicki et al., not only quantifies up to 3000 MFWs but also tests particular strata of MFWs, attempting to identify if Delta's success may be specific to a particular frequency rank. On the basis of the results obtained, Rybicki et al. recommend quantifying the first 3000 MFWs (Rybicki and Eder).

Once Delta's efficacy had been conclusively demonstrated, a number of scholars began to examine the possibility of extending it to other CLS problems. Rybicki et al. attempts to generalise Delta's functionality by applying it to other languages, attaining high levels of results in French, German, Hungarian and Italian corpora but poorer results for Latin and Polish (Rybicki and Eder). Rybicki et al. and Forsyth apply Delta to translated texts, in an attempt to identify whether the stylistic signal of the author or translator predominates. Both find that the signal of the original author is more powerful, but the presence of different translators can be identified by comparing two different translations of the same author's works (Forsyth and Lam; Rybicki and Heydel). Through use of bootstrap consensus trees and network analysis, which involves the representation of texts and the relationships between them as discrete entities (Eder, "Visualization in stylometry: Cluster analysis using networks"), Changsoo Lee, demonstrates that the further two languages are apart linguistically, the more likely it is that the translator's writing style will exert itself in comparison to that of the author (C. Lee). An example of how bootstrap consensus trees work in practice can be seen in Appendix B.

Apart from methods conducted solely through use of the Delta method, we see Alexis Antonia et al. utilise word-types identified through use of Burrows' Zeta method as authorial markers in order to investigate whether the presence of larger n-grams as opposed to individual words in specific chunks of text are more likely to correctly attribute authorship. Antonia et al. find that the efficacy of the variable depends on the corpus under consideration (Antonia et al.). This conclusion, that the optimal parameters and measures vary between corpora, seems to be confirmed by studies such as Enrico Tuccinardi, who demonstrates that character grams are more suitable in shorter documents (Tuccinardi) and Lisa Pearl et al.'s analysis of idiolect in epistolary literature, which allows for the weighting of some features as being more important than others (Pearl et al.). These findings culminate in the developing tendency within CLS towards electicism and the application of a diversity of methods applied to a similarly diverse set of parameters, whether discriminative words, word lengths, character-based frequency analyses, POS tags or measures for vocabulary richness, to which vector space representation, PCA, hierarchical clustering, SVM, random forests, k-nearest neighbours, Delta or rolling Delta may be applied (Gladwin et al.; Hou and Jiang; Saccenti and Tenori; Sayoud).

The conducting of multi-faceted statistical investigations such as these have been further aided by the development and dissemination of Eder et al.'s stylo package for the R programming language, which allows for the implementation of almost every major CLS method which has been proposed within the history of the discipline (Eder et al.). Stylo also provides a guided user interface (GUI) so that the construction of a relatively involved CLS pipeline is straightforward matter even for any would-be CLS critic with no prior statistical or programming experience. We see the benefits of stylo's dissemination in a number of recent CLS papers (Ilsemann, "Forensic stylometry"; Oakes). A recommendation to adopt a diversity of tactics and systematically compare the results has therefore become more or less a *sine qua non* of most CLS approaches in recent years (Jockers and Witten; Kocher and Savoy).

1.3.2 Literary Theory

The basic positions we confront in engaging these debates concerning the supposed incompatibility of CLS within literary criticism will by this stage in this chapter be familiar and the argument that CLS is both overly generalising and insufficiently reflexive as a form of scholarly inquiry, remains a point of attack which is frequently revived (Gooding). However, we have not yet considered CLS scholars who have made a virtue of this charge to a certain extent, as we see in the literary criticism of Franco Moretti. It should be noted that Moretti's major works such as *Atlas of the European Novel* (1998), *Graphs, Maps, Trees* (2005), *Distant Reading* (2005) and *The Bourgeois* (2013) are not substantively computational or statistical in their approaches, but rather use maps, spreadsheets and diagrams in order to illustrate what are often quite traditional literary-critical hypotheses, holding out the possibility that literary criticism might aspire to the ambition and scope of quantitative sociology (Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History* 4–30; Moretti, *Distant Reading* 67). Moretti's most notorious argument, that the development of industrial capitalism in nineteenth century Europe (Moretti, *Distant Reading* 16–18; Moretti, *The Bourgeois: Between History and Literature* 14–21) paves the way for the emergence of modernist literature, is not in and of itself a controversial one. As we will see in the fourth chapter of this thesis, this axiom more or less undergirds a significant amount of literary criticism conducted from a Marxian perspective. Moretti's reception has more to do with what is perceived as his method's apologia for the literary-critical school referred to as the world literary system as it has been developed by Pascale Casanova (Cleary, "The World Literary System: Atlas and Epitaph"). Criticism of this school has been trenchant from post-colonial scholars such as Emily Apter and Christopher Prendergast on the basis of its tendency towards national chauvinism, imperialist logic and uncritical handling of the relationship between modernisation and the canonisation of literature (Prendergast; Apter 42–58). However, the publication of Moretti's writings, and responses to them, in pre-eminent venues such as *New Left Review* and *n+1* (Allison, Sarah et al.; Moretti, "The Roads to Rome: *Literary Studies, Hermeneutics, Quantification*") has the consequence that these criticisms have a tendency to assume the shape of criticisms of CLS in general, despite the lack of actual quantification in Moretti's work. The fact that the specificities of Moretti's work in the context of the Stanford Lit Lab, which consist of straightforward frequency analyses of POS tags or word frequencies in a manner which is far less provocatively post-political (Algee-Hewitt et al.; Allison, Sarah et al.; The Stanford Lit Lab) have been taken up for negative criticism to a significantly lesser extent attests to the fact that it is Moretti's more traditional literary-critical work which can be criticised on the basis of its Eurocentricity.

Critics who dispute CLS scholars' dependence on reductive or categorical reasoning at this time begin to advocate instead for more exploratory or interpretative approaches (Escobar; Sinclair) and we might consider Steven Ramsay, Joanna Drucker, Bethany Nowviskie and Jerome McGann symptomatic of this tendency, given their proposals that humanities computing reconfigure itself as a synthesis of theory, statistics and aesthetics. In seeking to locate a common ground between the works of these critics, we might identify their joint rejection of ground truth. Any reductive striving towards 'accuracy' is rejected, in favour of a focus on a generative or procedural critical project which may emerge from the transformation of texts, according to the notion that deformance, re-mediation, translation and misprision form crucial parts of the critical enterprise (Drucker; Ramsay x; Rockwell). The difficulty in considering the work of these critics within the context of CLS is that, even though they may provide novel and engaging philosophical insights, they do not engage to a significant extent with the actuality of statistical approaches as they are deployed within CLS and it is as a result impossible, on the basis of their writings, to arrive at practical steps towards the implementation of a provisional or exploratory CLS.

Taylor Arnold et al.'s work, which considers the ways in which automated visualisation techniques may allow critics to look at large corpora before proceeding with their analyses, in its simultaneous consideration of both Barthes and the functionality of machine learning, seems to be the closest the field has yet come to combining the actual mechanics of computing with theoretical criticism (Arnold and Tilton). Ramsay's account of stylometry is by contrast limited to the close reading of frequency tables and measurements for vocabulary richness, rather than an engagement with the ends to which these methods have been put in the context of PCA, or indeed other multivariate approaches (Ramsay 70–73). There is also a tendency at work here to overlook the changing nature of CLS over time. While in its early days CLS scholars may well have had a propensity to overstate the significance of their results, by the 2000's we can see that the promises to re-found literary criticism around scientific rigour in order to exorcise the spectre of post-structuralism have given way to comparisons with endeavours such as sociology, economics or state planning all of which have long histories of applying statistics in critical and reflective ways. Burrows for example asserts that as it would be an impossibility for a demographer to identify 'pure' instances of the social phenomenon they aim to quantify, whether class, race or gender, the use of spectra or 'fuzzy logic' becomes essential and pragmatic points of departure such as this go a long way towards rejecting the caricature of a reductive or methodologically unsophisticated CLS which sceptics identify as operating within the discipline (John Burrows, "Rho-grams and rho-sets: Significant links in the web of words" 725). How much the field of CLS can be said to have advanced in this regard can also be seen in the changing nature of the anti-CLS articles which now circulate. While Nan Z. Da's criticism of CLS is partly inhibited by its aim to de-legitimise the quantification of literature in general, it still represents a change in criticisms of CLS, in that it argues that methods are widely misunderstood or not implemented properly. Implicit within Da's analysis then, is the notion that the field could be improved on these bases (Da). Katharine Bode, in a response to Da's article, also notes this distinction, as well as the greater degree of care which needs to be taken in critiquing CLS on the basis of its scientism, given the pivot from 'merely' empirical approaches to a greater emphasis on the subjectivity or uncertainty within for example, the modelling of machine learning outputs (Bode, *Computational Literary Studies: Participant Forum Responses, Day 2*; Bode, *Computational Literary Studies: Participant Forum Responses, Day 3*; Bode, *Computational Literary Studies: Participant Forum Responses*) and we consider these in the next section.

1.3.3 Machine learning

It has already been suggested that machine learning is crucial in allowing for a proliferation of approaches within CLS. These methods are noteworthy primarily from the point of view of the extremely high rates of success they have had in attributing authorship. In Graeme Hirst et al.'s application of SVM to syntactical bigrams, which attains an average of 90% classification accuracy in differentiating two authors from one another (Hirst and Feiguina). Eder also carries out two benchmarking studies which aims to investigate the best means of optimising machine learning techniques, finding that bags of words far exceed sequential samples (Eder, "Does size matter? Authorship attribution, small samples, big problem") and that NSC, SVM and Eder's Delta with 20 cross-fold validation exceed *k*-NN, Naive Bayes and Delta. Eder concludes that most frequent bigrams, including punctuation, seem to represent the optimal parameter (Eder, "Taking Stylometry to the Limits- Benchmark Study on 5,281 Texts from "Patrologia Latina"").

These machine learning studies are accounted for here for three reasons. Firstly, they represent a significant juncture within the history of CLS and potentially indicate the direction of the field in the future. Secondly, the incorporation of machine learning will not only represent a technical change, but potentially also a conceptual one which may even actualise the attempted return to qualitative readings we have seen consistently in our history of the development of the Delta method. We see this in Jack Elliott's application of an unsupervised machine learning method originally used to model gene expression data. Elliott finds this method to be highly effective in discriminating texts on the basis of authorship, even in the highly industrialised context of romantic publishing, but finds that the most discriminative word-types are medium frequency content words (Elliott). Thirdly, the primary aim of Ted Underwood's two books, *Why Literary Periods Mattered* (2013) and *Distant Horizons* (2019) is to use machine learning to reconceive literary epochs in less static terms, demonstrating that the distinction which is held to exist between nineteenth and twentieth century literature might be more productively viewed as a more long-term transition. Underwood aims to prove this, not only via computational literary criticism but also by democratising the means through which these analyses may be replicated, adapted or undertaken as in his ongoing collaborative work with the digital repository HathiTrust, which are, at time of writing, continuing to expand the number of digitised books available from its website for prospective CLS scholars' usage. In the context of a project which aimed to classify page-level data into one of three categories, either prose, poetry or drama, Underwood demonstrates how the two paradigms of knowledge production held to be in opposition for almost the entirety of CLS' history, the statistical and literary aspects, may be synthesised. Underwood notes that as literary critics do not understand genre empirically, but rather socially, it therefore makes no sense to enforce a rigid either/or classification, but rather an approach based on a spectrum, is much more appropriate. Approaches arising from the field of machine learning, with its capacity to score goodness of fit as a figure somewhere between zero and one, zero representing not certain and one representing total certainty, are uniquely suited. A further safeguard against empirical reductionism is erected by cross-validating the obtained results with human judgement, specifically a group of five readers who, through use of a GUI purpose-built for the project, were recruited in order to classify literary data page by page. Through the labour exerted by these readers, who labelled all pages in 414 books, training data for the project was obtained, which was instrumental in the algorithm attaining an agreement rate of 94.5% in identifying prose as opposed to poetry, fiction as opposed to nonfiction and body text as opposed to paratext. The statistical model which was constructed on the basis of this training data was found to be less accurate than human judgment by a margin of just 0.9%. In this way, Underwood's utilisation of machine learning points to the capacity of CLS to fuse conceptual ambiguity or shades of difference within an empirical approach (Underwood, *Understanding Genre in a Collection of a Million Volumes, Interim Report* 8–12).

1.4 Conclusion

In providing a history of the development of CLS, this chapter has demonstrated that from an early stage in CLS history, the frequencies of an undifferentiated selection of high-frequency word types were highly effective in identifying texts which were similar in style as likely to have been written by the same author. However, CLS scholars aimed to challenge the predominance of post-structural theories of authorship and as a result, CLS was from its inception subject to robust criticism from a cohort of literary critics who were more invested in theoretical readings, who charged CLS critics as operating within a politically reactionary and reductive form of knowledge production. In response, CLS cleaved from an early stage in its history to organic theories of authorship and a focus on unconsciously deployed formal features within the work. The initial breakthrough regarding the efficacy of highly frequent word types is consequently elided for a significant period of time in favour of focuses on the individual contributions of particular words or word types insofar as these can be re-integrated within a traditional or qualitative literary-critical reading. This remains the case even after Burrows develops the Delta method which subsequent CLS scholars render in more sophisticated ways; these analyses are noteworthy for their focus on particular words and apparent reluctance to move into higher and higher frequency strata. It is not until scholars such as Eder and Rybicki enact a sequence of benchmark analyses that the superiority of quantifying thousands of MFWs is rendered irrefutable, as well as the development of highly effective unsupervised machine learning techniques optimised for large datasets with thousands of parameters, within which manual intervention would become impractical or inefficient. The most fundamental change which this chapter has charted can be identified in the potential offered by the application of statistical methods to large CLS datasets which have become more freely available for public use due to the lower technological costs associated with producing them, finally rendering an historical CLS possible. Having documented the history of the Delta method's emergence, optimisation and popularisation and how it significantly enhances the prospects of a CLS rooted in historiography, this thesis will next outline what role the Delta method and machine learning played in analysing its own object of study.

Chapter 2

Methodology

This chapter will document the methodological aspects of this thesis' quantitative approach and will do so by outlining sequentially the process through which the analysis was enacted. This chapter begins where the previous chapter left off, taking the established statistical methods as they have been described and operationalised within CLS and applying them in our approaches to parametrisation, analysis and visualisation. This is done in order to identify sudden changes in the literary historical record, which we refer to as 'breaks.'

This chapter's adopted method can be divided into two parts. The first outlines the process through which the corpus was constructed and the series of decisions which eventually led to the use of the HathiTrust Research Centres' collection of word frequencies spanning the two hundred and twenty two year period from 1700 to 1922. The second will describe how the form in which the data appears necessitated its de-duplication as well as the particular statistical approaches which were ultimately adopted.

It should be noted that not every single procedure which was performed upon the data in the pre-processing, modelling and analytical stages will be accounted for in this chapter. For those who may be seeking to replicate the results obtained and outlined below in full for their own purposes, it will be necessary to draw both from what appears below as well as the individual R scripts which appear in the appendices to this thesis which have been annotated specifically for this purpose.

2.1 Corpus Composition

The first stage of the process involved compiling the corpus. In order to formulate a means of quantifying literary influence diachronically, a sufficiently large sample size of texts spread over an extended period of time would be required. In order for this range of texts to qualify as a legitimate object of study from a diachronic perspective, it would be necessary that the corpus contain contrasting numbers of texts from previous centuries, as well as popular and non-canonical texts drawn from a variety of national traditions, ideally with a gender parity between texts. This was alighted on as the best means of gesturing towards the orientations of literary studies at the present time, moving beyond books written and published in English-speaking countries in order to develop an insight into literary history on an international scale. A sufficiently large sample of texts which meets requirements from the point of view of national as well as gender parity could work together in collectively composing some secure points of comparison and perhaps also a degree of specificity regarding the formal and stylistic dispositions of particular epochs and the distinct rates of transmission attached to each.

It was anticipated that in satisfying them, a corpus would have to be compiled manually. This would be preferable to accepting the terms by which another project may have opted to make a corpus available online with an orientation towards a particular theme or collection, which can only very rarely provide a reliable portrait of publishing activity of a particular period. It was anticipated that the volunteer-led digital archive Project Gutenberg (PG) would be an ideal source in fulfilling all these requirements. PG's archive consists of more than 58,000 works in an open, plain text format which is ideal for computational analysis (McArthur et al.) and it is for this reason that PG's holdings are frequently depended upon as textual datasets within CLS (Guo et al.; Khmelev and Tweedie). In addition to this, PG's volunteers abide by a sophisticated workflow in digitising their holdings, which involves the continual revision of transcripts in a series of stages (Richardson and English). This allows PG to form a decisive contrast with other online sources for literature such as The Internet Archive, which automates the production of texts via Optical Character Recognition (OCR) software. Their output is significantly less accurate as a result (Richardson and English). However, it turned out that the goals motivating the composition of a corpus are more easily devised than executed. PG's browse experience is quite poor and the website aims to prevent users from obtaining large numbers of texts at once. Each of PG's file-ids are unique to PG and do not reflect any other metadata convention such as ISBN or OCLC. It also became clear that PG tends to reproduce many of the already extant biases of literary canonicity; one need only examine PG's 'top 100' section in order to perceive that the authors PG has been most successful in disseminating, such as Jane Austen, Mary Shelley and Jonathan Swift, are already very well-known. It also turned out that PG's holdings were highly Anglocentric and very temporally weighted. As an archive, PG is far better equipped to cater for texts written and published in the nineteenth century than any other. This is itself indicative of problems within CLS more generally. Until the advent of industrialised production publication was far rarer; the uneven concentration of texts from the more recent past in many ways reflects this (Caruana-Galizia 447). In addition, translations of modern literature were very rare in PG, and certainly not available at a scale which would provide the means for this project to incorporate a non-English literary tradition to any significant extent. In accounting for PG's national and temporal biases, we might note that PG is an American website and only novels which have been published before 1924 and have entered into the public domain in the United States can be held online by PG without violating US copyright law. PG does have Canadian and Australian derivates — presumably operating in these jurisdictions in order to benefit from more lenient copyright regimes — but as before, those texts which are present also skew towards, rather than away from, the established canon to incorporate authors such as Evelyn Waugh, Virginia Woolf and Sinclair Lewis. In summary, PG's status as a volunteer-led digital archive, as opposed to one allied to any particular institution or library, means that it does not abide by any impulse towards the preservation or digitisation of materials beyond the interests of their individual volunteers. This might account for the apparent prioritisation of texts which are already popular as opposed to more obscure ones. Finally, PG introduces difficulties of a more pragmatic nature; the only search functions which PG provides are based on keyword searches or a browse function which lists works alphabetically by author surname. It is therefore impossible to search by the fields in which we are most invested: author nationality, date of composition or publication. When these metadata are present, they can only be found within the documents themselves and are therefore inaccessible on a broad scale.

Some of the shortcomings associated with PG as a resource then, occasioned further reflection on what it would mean to incorporate or analyse texts which have fallen victim to what Moretti refers to as the 'great slaughterhouse of literature.' Within accounts of modernist writing for instance, it is obvious that James Joyce or Virginia Woolf would receive greater amounts of attention than Wyndham Lewis or Djuna Barnes, but whether these latter two could be said to be exemplars of the great unread is certainly debatable, especially when the recovery of voices held to be liminal or formerly obscured characterises much of contemporary literary criticism (Mao and Walkowitz, "The New Modernist Studies" 737–38). When Moretti refers to the slaughterhouse of literature in the context of CLS, Moretti seems to be envisioning the resurgence of texts which have never, to any significant extent, been accounted for within the corpus of literary criticism. The great slaughterhouse of literature might therefore be better grasped as a sort of absolute, or a utopian concept which gestures towards the potential offered by CLS in general terms, rather than an actual descriptive category which might be usefully attached to particular authors. This line of argument would be further attested to by Moretti's presentation of Arthur Conan Doyle, Bram Stoker and Charles Dickens as authors who are often omitted from histories of the novel (Moretti, *Atlas of the European Novel 1800 - 1900* 14; Moretti, *Distant Reading* 68–70). Once we attempt to operationalise the great slaughterhouse in the context of CLS and begin to grapple in a real sense with the availability of digital texts and the composition of the infrastructures which exist in order to facilitate their dissemination, some of the methodological tensions located within this concept then come into greater amounts of focus. We might also consider the nature of the commercial publishing infrastructure in a pre-digital age. As Lucien Febvre and Henri-Jean Martin note in their history of printed text, publishers have historically operated their printing operations on the basis of texts which could be relied upon to obtain a profit. As they write:

> We should not therefore be surprised to find that the immediate effect of printing was merely to further increase the circulation of those works which had already enjoyed success in manuscript, and often to consign other less popular texts to oblivion. By multiplying books by the hundred and then thousand, the press achieved both increased volume and at the same time more rigorous selection. If we keep that fact in mind we shall understand better the nature of the printing industry in the fifteenth century. (Febvre and Martin 249).

It is therefore not as though the project of ascribing canonicity takes place exclusively within modern academia, exterior to the available materials; rather the materials that we have are constitutive of the project of ascribing canonicity itself. As an exemplar, we might consider the number of barriers which a seventeenth century French play written by a woman would be required to circumvent in order for it to be transcribed and disseminated widely on the internet in an open format; the notion of a CLS project which can bring greater amounts of attention to obscure works would seem to be a kind of category error. It is therefore the case that if any meaningful challenge to the canon is to take place within the context of CLS in general, or computational literary projects in particular, the project must incorporate some digitisation of otherwise inaccessible materials to be

Table 2.1: HathiTrust dataset contents

Genre	Quantity
Fiction	101948
Poetry	58724
Drama	17709

held or presented on an online infrastructure optimised for the further dissemination and usage of said materials.

2.2 HathiTrust

While the corpus composition of this project was ongoing, the CLS scholar Ted Underwood published his second book, *Distant Horizons* (2019). This text spurred some further examination of the potential offered by the HathiTrust Research Centre (HTRC), a body which hosts the HathiTrust Digital Library (HTDL), a large-scale repository of digital content obtained from the holdings of local and research libraries which have been digitised by Google and the Internet Archive. Combining all these sources or infrastructures together as HathiTrust have done, culminates in about 4.8 million volumes which are in the public domain. HTDL provide a sample of this archive online in machine readable formats and most importantly in the context of this project, in the form of the 'Word Frequencies in English Language Literature 1700 - 1922' collection, an overview of the contents of which can be seen in Table 2.1 (Underwood et al.). As we see in Table 2.1, it is works of prose fiction that represents more than half of the texts in this dataset, poetry just over a third, and drama less than a tenth.

After some examination and testing of this resource, it was decided that this online dataset provided the best available means of undertaking a *longue-durée* quantitative analysis of literary production. In so doing, it was further confirmed that in sectors of the internet where reliable digital texts can be sourced such as PG and HathiTrust, the canonical is highly overrepresented. While initially this project set out with the aim of orientating itself relative to a number of approaches currently regnant within literary studies, such as the incorporation of a variety of national traditions as well as more popular forms, a trend which will be dealt with in this thesis' fourth chapter, from an inspection of the metadata attached to this collection of word frequencies the institutions from which the HathiTrust obtains its holdings are overwhelmingly concentrated within the United States, with the result that American and English works are prominent to a large extent. This can be seen in Tables 2.2, 2.3 and 2.4, which outlines the authors whose works occur most frequently in each of the three categories into which the collection is sorted. It should be noted that Tables 2.2, 2.3 and 2.4, were generated after the normalisation and de-duplication procedures outlined below have been carried out, but before mean word frequencies were calculated by volume and title.

It should also be noted that the most frequently occurring value in the 'author' column across all three datasets is, by a significant margin, a blank entry. The authors which are prominent in each of the three datasets are primarily canonical, male and English or American. The only exceptions to this are one French novelist and one woman in 2.2, one Norweigian, one Irish and one German in the drama dataset. From a qualitative inspection of the metadata, it can be confirmed that these tables are representative of the dataset. Popular and literary works written by English men predominate but Irish, French, German and American novelists are also present. Even with these shortcomings taken into account however, the HTDL's word frequencies dataset offer the significant advantage of immediate access to tens of thousands of literary works, far exceeding what would be feasible for a single individual working within the timeframe of this project to improve upon.

The three categories into which this collection of word frequencies have been sorted — poetry, prose and drama — is itself an outcome of Underwood's 'Understanding Genre in a Collection of a Million Volumes' project, described in the previous chapter. These categories were used because they are distinct enough to operate on a more secure basis than more specific generic categories which overlap to a greater extent, such as the gothic,

Table 2.2: Most frequently occurring authors (Fiction)

Author Name	Number of works
	1385
Balzac, Honoré de,	264
Braddon, M. E.	203
James, G. P. R.	198
Lytton, Edward Bulwer Lytton,	185
Oliphant,	169
Cooper, James Fenimore,	160
Scott, Walter,	158
Trollope, Anthony,	158
Dickens, Charles,	136

Table 2.3: Most frequently occurring authors (Poetry)

Author Name	Number of works
	1360
Browning, Robert,	81
Tennyson, Alfred Tennyson,	59
Byron, George Gordon Byron,	50
Longfellow, Henry Wadsworth,	47
Scott, Walter,	46
Morris, William,	43
Moore, Thomas,	28
Riley, James Whitcomb,	26
Young, Edward,	25

Table 2.4: Most frequently occurring authors (Drama)

Author Name	Number of works
	317
Pinero, ArthurWing,	46
Shaw, Bernard,	45
Jones, HenryArthur,	44
Ibsen, Henrik,	37
Hare, WalterBen,	32
Knowles, JamesSheridan,	25
Galsworthy, John,	23
Maugham, W.Somerset	21
Sudermann, Hermann,	20

science-fiction or fantasy (Underwood, *Understanding Genre in a Collection of a Million Volumes, Interim Report* 2–3). The motivating impulse behind Underwood's genre project was to provide a replicable means of analysing large, heterogenous digital collections with no, very little, or unreliable metadata, as is the case for the collections which Google's digitisation initiatives of the past decades have left in the hands of a number of institutions across the world, both public and private. Analysing large resources such as these therefore present fairly extensive difficulties; HathiTrust for example, has more than 13 million volumes, around 276 million pages; it would take an enormous amount of time to develop accurate metadata for all these works manually. Underwood's report underlines this, but also proposes that collections such as these provide the best available means of elevating the great slaughterhouse of literature from an idealistic proposition into methodological actuality, especially as the cultural output of the twentieth century begins, printing and publishing output increases exponentially and legal prohibitions begin to seriously impede our work as CLS scholars (Underwood, *Understanding Genre in a Collection of a Million Volumes, Interim Report* 4). Underwood presents a solution to this problem by framing it as a classification task, analysing OCR data obtained from every page in 854,476 volumes in the Python programming environment. 1062 page-level features were identified as parameters; 1036 of these were words and 26 were structural features, such as margin size, number of pages in the volume and total number of words on the page (Underwood, *Understanding Genre in a Collection of a Million Volumes, Interim Report* 21). Analyses were carried out on the level of the individual page out of necessity; as was previously mentioned, the metadata is scarce and multiple genres are also often contained within single volumes. There may, for example, be a single volume which contains both poetry and drama. To further complicate matters, these volumes may have prose introductions or indexes (Underwood, *Understanding Genre in a Collection of a Million Volumes, Interim Report* 6). By identifying the relative frequencies of the words and the presence, absence or continuous values attached to each structural feature, it is possible to regress these word frequencies against the categorical variable of genre;

Table 2.5: Certainty of correct classification (Fiction)

x
Min. :0.6800
1st Qu.:0.8336
Median :0.8467
Mean :0.8386
3rd Qu.:0.8533
Max. :0.9000

Table 2.6: Certainty of correct classification (Poetry)

x
Min. :0.7800
1st Qu.:0.8158
Median :0.8206
Mean :0.8229
3rd Qu.:0.8274
Max. :0.9233

to identify for example, the likelihood that the text under analysis is a prose text. This calculated likelihood is incorporated into the metadata associated with each volume in order to allow CLS scholars to filter the available material according to the requirements of their research, allowing for greater or lower amounts of rigour. The statistic can be taken as indicative of the relative certainty that 80% or more of the text's pages have been correctly classified. At first we explored the possibility of incorporating this statistic into our analysis as a form of error modelling, but it was found that at the scale at which we are most interested - fluctuations of stylistic difference on an annual basis - uncertainty remains relatively constant, as can be seen in Tables 2.5, 2.6 and 2.7.

Once the data had been de-duplicated and merged by title, certainty never fell below 68% and the overwhelming majority of values are over 80%. Underwood demonstrates that restricting one's analysis to a threshold above 50% certainty obtains only a minimal gain in precision, with the result that one is dimissing a disproportionate number of accurately

Table 2.7: Certainty of correct classification (Drama)

x
Min. :0.7567
1st Qu.:0.8067
Median :0.8169
Mean :0.8174
3rd Qu.:0.8273
Max. :0.8700

classified texts (Underwood, *Understanding Genre in a Collection of a Million Volumes, Interim Report* 33). Given both Underwood's own judgement and the consistency of these values displayed in Tables 2.5, 2.6 and 2.7, it was decided that there was only limited value to be gained by incorporating relative amounts of certainty or uncertainty into our analysis and we would not, therefore, be making use of these statistics.

2.3 Data Cleaning

In order to facilitate CLS scholars' making use of this resource to a greater extent, HTDL have also uploaded aggregated summary files which includes the 10,000 MFWs of every genre by year. This would probably have been this thesis' starting point, but as the HTDL website notes, duplicate texts are included within this aggregation (Underwood et al.). Since this research project attempts to enact a diachronic analysis and we are interested in characterising the stylistic portrait of every year as accurately as possible within the limits of the metadata provided, it was decided that the original .tsv files containing the word frequencies for each individual volume would be our starting point instead. An example of how the word frequency data was structured appears in Table 2.8. The particular way in which this data appears had significant consequences for this study's approach. The fact that the texts appear in the form of single words made any modelling of broader textual entities, such as phrases, *n*-grams, sentences or topics impossible and rendered Delta, a multivariate method with a robust capacity to return

Table 2.8: Fifteen rows of a randomly selected .tsv file

Word	Frequency
.	251336
—	150209
,	143299
i	93619
the	52736
1	48718
3	41931
v.	40230
2	38774
of	36908
iv.	36792
to	36049
and	34516
ii	34277
iii	31015

accurate results using relative frequencies of individual words, the most viable means of proceeding. It should be noted that Underwood provides a number of .csv files which can be used in order to correct some commonly occurring OCR errors, especially in the older texts, with the disclaimer that they are not universally accurate and depend on context as well as the interest or needs of the researcher.

In this particular research project we wanted to do as little normalising of the data as possible so we could view our results in as unmediated a form as possible. We did not want, for example, to modernise the antiquated usage of words so much as to have them form part of our study. We therefore did not want to remove personal or place names, remove abbreviated usage ('abandon'd' as opposed to 'abandoned') or fuse words which appear in twos ('every where' as opposed to 'everywhere'). All of these seemed to be examples of changes in usage that we actually interested in tracing. We did however, wish to correct incorrect spelling, presumably because these would be a product of OCR errors ('califomia' as opposed to 'california') and remove roman numerals, which are far more often indicative of chapter or section headings rather than content. We also wanted

Table 2.9: Sample of drama metadata

htid	author	title	enumcron
hvd.hxvf6w	Shakespeare, William,	The works of Mr. William Shakespear	v.7
pst.000055157780	Rowe, Nicholas,	Rowe's plays	v.2
mdp.39015021040012	Dryden, John,	The dramatick works of John Dryden, esq;	v.6
mdp.39015021040020	Dryden, John,	The dramatick works of John Dryden, esq;	v.5
mdp.39015021040046	Dryden, John,	The dramatick works of John Dryden, esq;	v.2

to standardise spelling where this was necessary ('apologise' as opposed to 'apologize').

There were 178,381 of these .tsv files, one for every volume in each of these three genres and together this comprises about 12.5 GB of data. These files are contained in zipped folders grouped together on the HathiTrust website both by genre and publication date. These folders were downloaded, unzipped and placed in one of three relevant folders in our working directory, 'fiction,' 'poetry' or 'drama.' Each .tsv file for associated metadata was downloaded separately.

An example of four rows of the drama metadata can be seen in Table 2.9. This metadata provided the primary means through which each of these three datasets were engaged. Each value in the first column labelled 'htid' corresponds to a particular .tsv filename and it is through this that we join each table of frequencies to the metadata, allowing us to tie each table of frequencies to a particular author, title and date of publication. Through this we de-duplicate and otherwise begin to normalise the corpus. Before doing this though, we plotted the date of publication data on a series of histograms in order to identify just how even our spread across the two-hundred and twenty-two year period actually was. The results for each dataset appear in *Figs.* 2.1, 2.2 and 2.3.

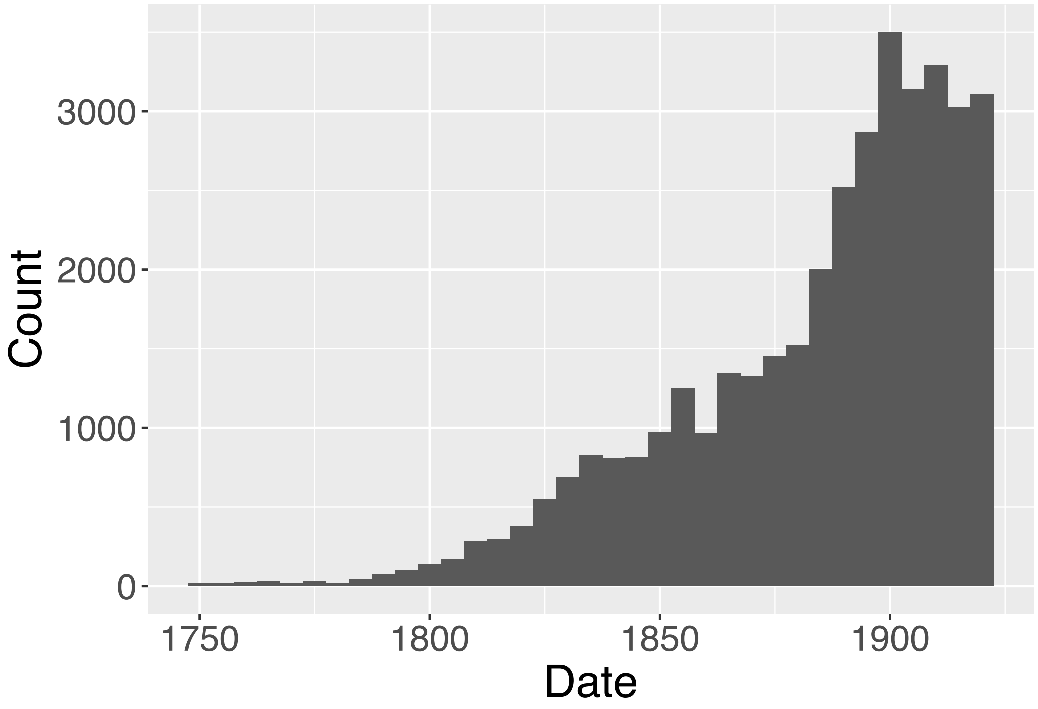


Figure 2.1: Distribution of publication date (Fiction)

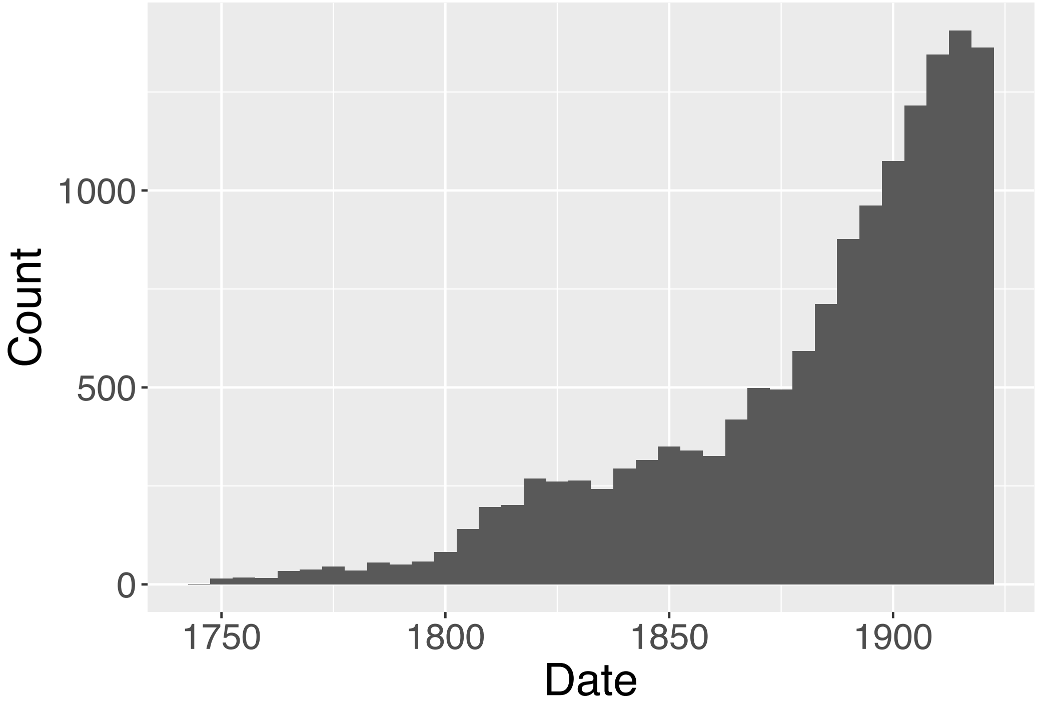


Figure 2.2: Date of publication date (Poetry)

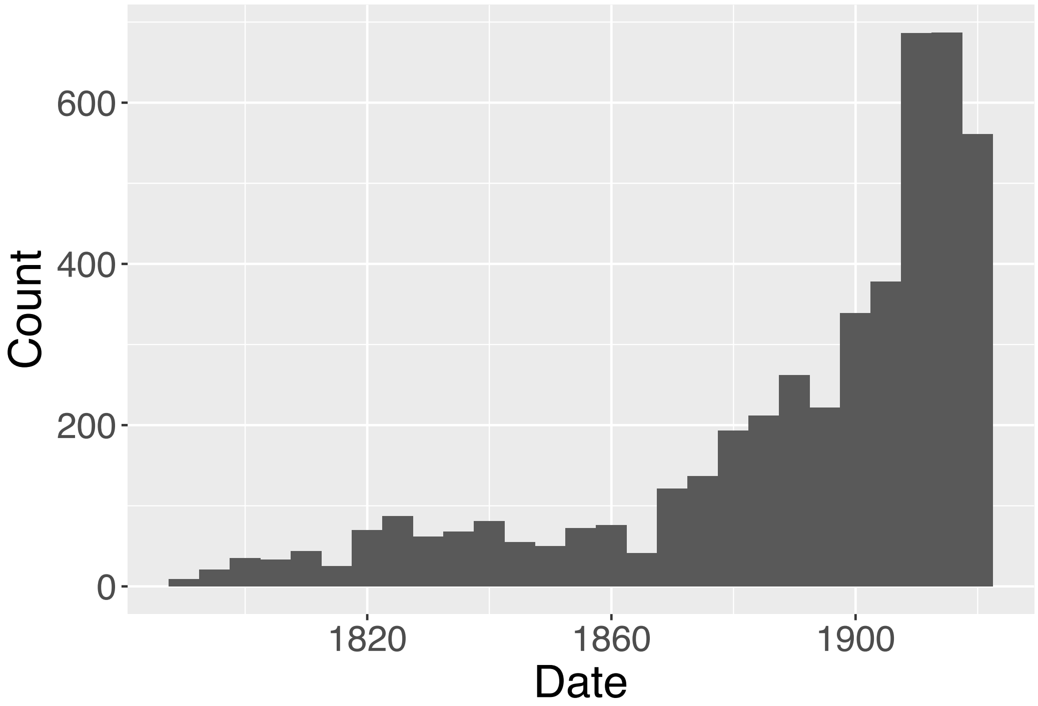


Figure 2.3: Date of publication date (Drama)

As can be seen in *Figs.* 2.1, 2.2 and 2.3, the overwhelming majority of texts contained in the dataset are concentrated towards its more recent end, i.e., we have far more texts from the nineteenth century than we do the proceeding eighteenth. Though we have far more texts than we could attain through any other means, HathiTrust's database has still not allowed us to escape the problem outlined in this chapter's opening section: nineteenth century texts are far more easily accessed in digital format than any century before or after. This is due to the fact that books published before 1800 are more likely to be held within the more valuable or fragile section of any given libraries' holdings and are therefore less likely to be handed over in the event of a digitisation initiative to the extent that other, more recent publications are. HathiTrust therefore warns that the data should probably not be used in order to make any extensive generalisations about literature published before 1750 (Underwood et al.). Rather than taking this disclaimer as an injunction, we decided to concentrate on the point at which gaps in publication stopped appearing. Identifying both the beginning and end point of this continuous line of publication necessitated some de-duplication and normalising of the data.

2.4 De-duplication and joining the datasets

Underwood identifies results obtained from an analysis of the drama dataset as being prone to distortion due to the degree to which duplicate texts appear in the form of re-prints of Shakespeare or other Elizabethan playwrights, but argues that duplicates effect the drama and poetry corpora 'only to a moderate extent.' Underwood also explains that this corpus was compared to smaller collections of first editions and that the differences between the two were 'not particularly shocking' (Underwood, *Understanding Genre in a Collection of a Million Volumes, Interim Report* 37). Underwood however provides no specific information or benchmarks regarding the number or effects of re-prints in the fiction or poetry corpus. Though Underwood does provide some de-duplication techniques on his GitHub page (Underwood, *DataMunging:rulesets at master · tedunderwood:DataMunging · GitHub*), these are written in the form of Python scripts which were constructed in order to pre-process the data for specific machine learning tasks and they are therefore not relevant to us in this instance. It was therefore decided to de-duplicate the texts manually and automate this task whenever possible. Taking this long option is not without its advantages; in addition to facilitating greater amounts of accuracy — automatic de-duplication would cause more inaccuracy than it would solve when the metadata is as irregular as it is in this instance — it also allows us to develop a greater sense of the corpus' overall composition. As can be seen in Table 2.9, there are a number of columns which may be used in order to de-duplicate the dataset. We began with a frequency analysis of the words contained in the 'title' column, where a number of collected works by heavily anthologised authors were identified, through an overabundance of titles containing words such as 'works,' 'novels' and 'Balzac.' The texts with 'works' and 'novels' in the title were overwhelmingly suggestive of the works of Dickens. The removal of these texts on the basis of the accompanying metadata was accomplished manually, until the frequency of these terms and others suggestive of them fell below ten, at which point it was intuitively assumed that it would be too granular a procedure to continue further and the trade-off in terms of manual work versus the gain in terms of accuracy would begin to decline. Anachronistic texts were also removed from the prose (Swift, Homer [sic], and Bede), poetry (Edmund Spenser, William Shakespeare, John Dryden) and drama corpora (Ben Jonson, Euripides, Sophocles). By running a frequency analysis on the 'subjects' column in the same way, we identified a number of texts which were likely to be non-fictive, based on the presence of words such as 'Gift books,' 'IsBiographical' and 'Conduct.' As we would wish to attenuate the influence of this genre in our dataset, it was decided that these texts would be removed also. It should be noted that this procedure probably did not lead to the total elimination of all non-fictive texts from our dataset; the 'subjects' column is among the least reliable columns and 72% of texts in the metadata have a blank entry in the cell. As before, once the frequency analysis identified repeating entities lower than ten, they were not pursued any further. Once the removal of anachronistic and the most obviously re-issued texts had

Table 2.10: Multi-volume versus single-volume works

	Multi-volume	Single-volume
Fiction	11110	26534
Poetry	839	13755
Drama	213	4414

Table 2.11: enumcron examples

Var1	Freq
BUH	1
?vol.?	1
. 1	1
(1826)	1
(1827)	1
(1828)	1
(1829)	1
(copy 1]	1
(pt.1-2)	1
(v.1-3)	1

been completed, the next step was to accommodate texts spread across multiple volumes. The number of multi-volume versus single volume works can be seen in Table 2.10.

As can be seen in Table 2.10, multi-volume works are far more likely to occur in the fiction corpus (30%) than in the poetry (6%) or drama corpus (5%). The relative presence of multi-volume works played no role in determining the procedure we enacted on them, which was identical in each case, but is highlighted more as a matter of interest. It was decided that the word frequencies for multi-volume works would be averaged across volume and that the date in which the first volume was published would be taken as the date of publication for each subsequent volume in any given series. Unfortunately, the volume column 'enumcron' is highly irregular, as can be seen in Table 2.11.

In the fiction corpus, there are 544 distinct entities in the volume column, very few of which can be identified automatically as representing a particular volume number, while poetry contains 215, and drama 100. Attempts were initially made to normalise these

Table 2.12: Data overview

	Total Works	Start	Finish	Total Years	No. of MFWs	Min. Text Size
Fiction	37644	1751	1922	171	8696	5266
Poetry	14594	1747	1922	175	4431	1507
Drama	4627	1791	1922	131	3751	1903

data points via finding and replacing particular combinations of characters, but after some experimentation it was found that the variable was too inconsistent to be standardised without manual intervention; there were too many instances which required individual judgement and deference to the accompanying metadata. A schema which could assist in classifying the entries 'v.1-3,' 'Third series' and 'ser.2' within the timespan of the project would be as likely to re-create as many mistakes as resolve them.

Once the 'enumcron' column had been standardised as outlined above, it was a far more straightforward matter to de-duplicate single volume works; the metadata was sorted in ascending order by date and titles which appeared more than once were removed. Multivolume works were removed by both title and volume number so as to increase the likelihood that we were removing actual duplicates from our dataset. Finally, at a later stage in our analysis, once we had identified our breaks in the historical record, we returned to the metadata in an attempt to identify particular works which may have been decisive in executing a given break. In doing so, more anachronistic texts were identified and removed.

As can be seen in Table 2.12, the de-duplication and removal of anachronistic works has reduced the number of works in our corpus by 68%. Fortunately, even with this attenuated number of works, the HathiTrust repository still far exceeds any other modern corpus of literature available online. As it obviates the necessity of constructing one manually, and spans approximately a century and a half, this corpus was ideal for a research project of this scope and subject.

2.5 Data Capture

It was then possible to load each .tsv file into the R workspace. R, as opposed to other ready-made text analysis tools, was used in this instance for the reason that it offers an industry standard programming environment for which a large community of users have developed code and libraries for others to use. As a result, R therefore offers the opportunity to employ a wide array of data manipulation, statistical analysis and visualisation tools, with a significant history of application within CLS (Feng 696).

As can be seen in Table 2.8, the frequency of each word in every text is represented in the form of its raw counts, meaning they are not normalised or expressed in terms of a percentage of the text's overall length as Delta requires. Therefore the second column of each text was summed and each word's raw count was expressed as a percentage of this sum. We then averaged these word frequencies three separate times, first by volume, title and date, then by title and date and finally, by date.

2.6 Most frequent words

As mentioned in the previous chapter, there is little consensus within the discourse regarding how many MFWs should be included within any given quantitative analysis, other than an increasing accumulation of evidence that the more the better. In pursuit of some kind of an objective benchmark for MFW usage in the context of this project, it was decided that we would obtain total word counts for every text in each dataset and adopt the third quartile of word counts as our MFW count. This approach not only has the advantage of providing an objective benchmark but it also allows us to move from high frequency words which are difficult to interpret from a qualitative standpoint out of context ('the,' 'and,' 'of') and more towards medium and low frequency words which are more relevant from the point of view of content ('bewilderment,' 'misgivings' 'account'). The number of MFWs we therefore adopted in each of our three datasets can be seen in Table 2.12.

Table 2.13: Relative frequencies of 5 MFWs

	a	aback	abandon	abandoned	abashed
1751	0.9172768	-0.1955380	0.02857518	0.3102666	1.6850156
1752	0.6923411	0.1074349	-0.49811503	-0.3874456	-2.3531625
1753	0.3809051	-1.0390224	0.11115178	-1.7381298	-0.8046574
1754	0.5309007	-1.0390224	-0.72914445	0.1892048	-2.3531625
1755	0.6809011	-1.0390224	-0.27856878	-1.5480914	-2.3531625
1756	0.5310893	-1.0390224	1.14630712	0.1035169	-2.3531625

As was also touched upon in the previous chapter, the lack of objective benchmarks for MFW usage finds its corollary in the lack of an ability to identify a precise point at which a text is too short for a Delta analysis, or at least a definitive benchmark for identifying the stage at which the efficacy of the Delta method begins to decline precipitously. Eder presents 5000 words as a minimum requirement (Eder, “Rolling stylometry”) but as we saw in Section 1.3.1, we would wish to avoid recommending unconditional text length in analyses when they seem highly contingent on the corpus under analysis (Jannidis et al.; Smith and Aldridge). We therefore excluded texts which were shorter than the first quartile of word lengths. This minimum value size is detailed in Table 2.12.

We inspected the resultant vector of MFWs for each dataset and removed words which were deemed unsuitable, such as numbers, punctuation or OCR errors, which appeared as words such as ’*R’ or similar. For every one of these word types which was removed we extracted an equal number of frequent words which appeared next in the frequency rank until we had a data frame which consisted of every MFW for every year contained in the dataset, with each year represented by a different row of MFW frequencies. A sample of the result can be seen in Table 2.13.

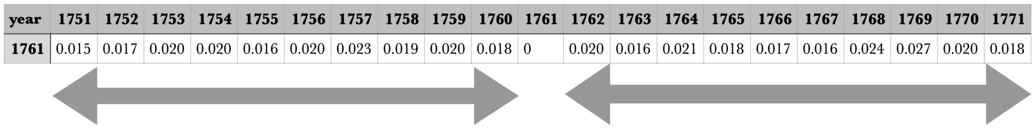
Table 2.14: Cosine distance between five years

	1751	1752	1753	1754	1755
1751	0.0000000	0.6223650	0.6654140	0.7022123	0.6252553
1752	0.6223650	0.0000000	0.6652436	0.6936948	0.6673178
1753	0.6654140	0.6652436	0.0000000	0.7115641	0.6286374
1754	0.7022123	0.6936948	0.7115641	0.0000000	0.6851975
1755	0.6252553	0.6673178	0.6286374	0.6851975	0.0000000

Cosine distance was then applied to this data object row by row, in line with Evert et al.'s findings that this is the most effective distance metric to apply in the context of a Delta analysis (Evert et al. 14). This distance figure was then divided by the number of MFWs, as is practiced in Delta. The result was a distance matrix, a sample of which can be seen in Table 2.14.

2.7 Identifying breaks

The first necessity in analysing our distance matrix is to identify a statistical baseline at which the presence or absence of a 'break' can be identified. This problem can be framed in terms of a deceptively straightforward question: at what scale does literature change? In answering this question, we may consider a number of different approaches which have been adopted within research projects initiated with a similar question in mind. We might adopt Moretti's 'twenty-five to thirty years' as a turnover for literary generations and take the opportunity the HTRC's word frequencies collection presents to test this concept (Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History* 21). In their temporally-based topic model of innovation and influence in the speeches and debates of the first parliament of the French revolution, Barron et al. adopt 36 speeches as a statistical baseline. Given that there are a total of 44,953 speeches in their corpus overall, this amounts to a moving window of effectively 0.08% of their data. If we were to adopt this percentage as a baseline for our own data, this would have us analysing it in fourteen year windows and perhaps contextualise our rate of change relative to how quickly spoken political discourse changes. However, the problem with this benchmark is that it is adopted more in order to reflect the number of speeches which are delivered in the National Constituent Assembly (NCA) over the course of two days. In this sense, there may be a certain degree of inappropriateness to generalising the benchmark, given Barron et al. apply it as an approximate measure of the political agenda in this particular context. Ultimately, we chose a third option and decided to investigate rates or extents of change within this particular dataset, without having a particular scale or approach imposed upon us. The best means of approaching this question would seem to be offered by the student *t*-test, an established means of assessing the presence of statistically significant differences between the mean values obtained from two numerical vectors. Though *t*-tests are an established method for problems of this kind they have the drawback of requiring a particular number of samples in order to function effectively; less than twenty samples is not regarded as being sufficient to secure a reliable outcome. The consequence of this for our analysis is that the first and last ten years contained within each of our corpora are passed over in being considered as enacting any kind of ‘break’ in any real sense; though they play their role in the analysis, there is not enough data on either side for their individual impact on the corpus to be assessed.

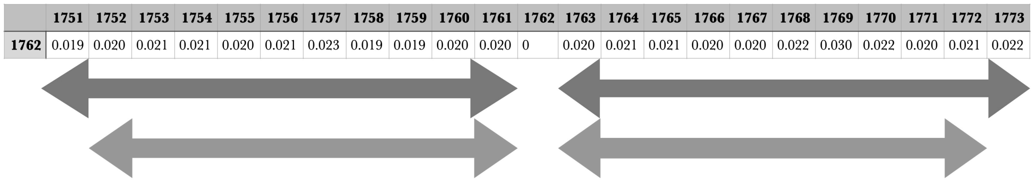


year	1751	1752	1753	1754	1755	1756	1757	1758	1759	1760	1761	1762	1763	1764	1765	1766	1767	1768	1769	1770	1771
1761	0.015	0.017	0.020	0.020	0.016	0.020	0.023	0.019	0.020	0.018	0	0.020	0.016	0.021	0.018	0.017	0.016	0.024	0.027	0.020	0.018

Figure 2.4: One t-test being applied

For example, the drama corpus begins in 1751 and ends in 1922. We therefore have to skip the first ten rows of our distance matrix and begin our analysis on the eleventh row, 1761, as we see in *Fig.* 2.4. We apply *t*-tests to identify the mean difference which exists between the vector of distances from 1751 to 1760 and compare this with the vector of distances between 1762 and 1771.

We see two *t*-tests being applied to one row in our distance matrix, in this instance a row representing the distances relating to the year 1762. In *Fig.* 2.5. The topmost arrow



	1751	1752	1753	1754	1755	1756	1757	1758	1759	1760	1761	1762	1763	1764	1765	1766	1767	1768	1769	1770	1771	1772	1773
1762	0.019	0.020	0.021	0.021	0.020	0.021	0.023	0.019	0.019	0.020	0.020	0	0.020	0.021	0.021	0.020	0.020	0.022	0.030	0.022	0.020	0.021	0.022

Figure 2.5: Two t-tests being being applied

represents the eleven value *t*-test, while the bottommost arrow represents the ten-value *t*-test. This is undertaken because once we enter onto this twelfth row, 1762, there is a berth of at least eleven years on either side so enacting multiple *t*-tests becomes possible. We therefore run a *t*-test at a scale of eleven years and then at a scale of ten years. We continue in this way, running a twelve, eleven and ten year analysis on the thirteenth row, a thirteen, twelve, eleven and ten year analysis on the fourteenth row, and so on, until we run an eighty-four to ten year analysis on the eighty-fifth row, which is the distance matrix's halfway point. At this stage, the number of years on the far side of the distance matrix begins to be eroded and the process begins to move in reverse. The result of every one of these *t*-tests provides us with a statistic known as a *p*-value. The established means of discerning the significance of a *t*-test is to identify whether or not the *p*-value is lower than 0.05, which would indicate that the probability that this result arose by chance is less than 5% (Crawley 92). As we are running a large number of *t*-tests at once, there is a risk that we have taken in a large number of false positives; this is one of the drawbacks associated with running multiple *t*-tests at once. We applied Yoav Benjamini and Yosef Hochberg's False Discovery Rate (FDR) algorithm to a vector which contained our *p*-values, thereby limiting the number of false positives which are reported as being significant, effectively employing a more robust benchmark for significance (Benjamini and Hochberg). Details about these *t*-tests such as how many were carried out, how many were identified as being significant, both before and after this false detection algorithm was applied, can be seen in Table 2.15.

Table 2.15: Statistically significant t-tests

Corpus	Tests	Significant Tests (FDR)	Rate of change	Start year	End year
Fiction	5853	3459	57	1751	1922
Poetry	6163	4333	50	1747	1922
Drama	3193	1604	38	1791	1922

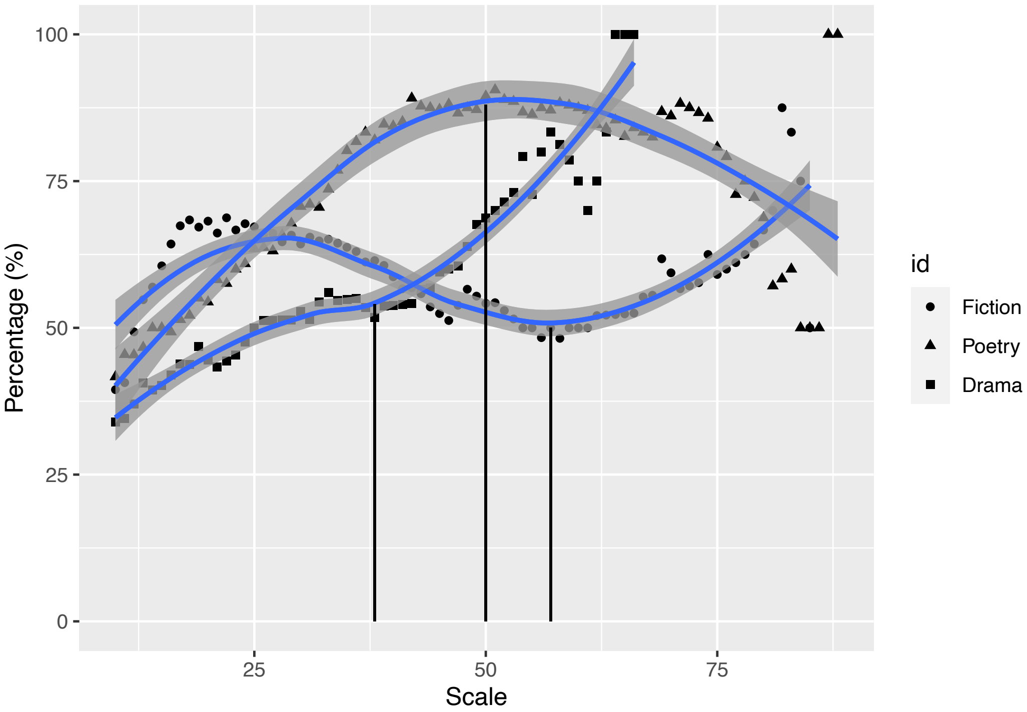


Figure 2.6: Percentage of significant t-tests for period of time

In order to identify which scales are the most significant, we modelled the aggregated results of the *t*-test against scale of time in each corpus. The result appears in *Fig* 2.6.

Each time a significant difference is identified, we subtract the two means from each other in order to calculate the difference. For each scale we summed these mean differences to create a cumulative figure which we then divided by the number of times the *t*-tests returned a statistically significant result overall; we call this statistic magnitude. If we model mean magnitude against scale in all three corpora, the result appears in *Fig* 2.7.

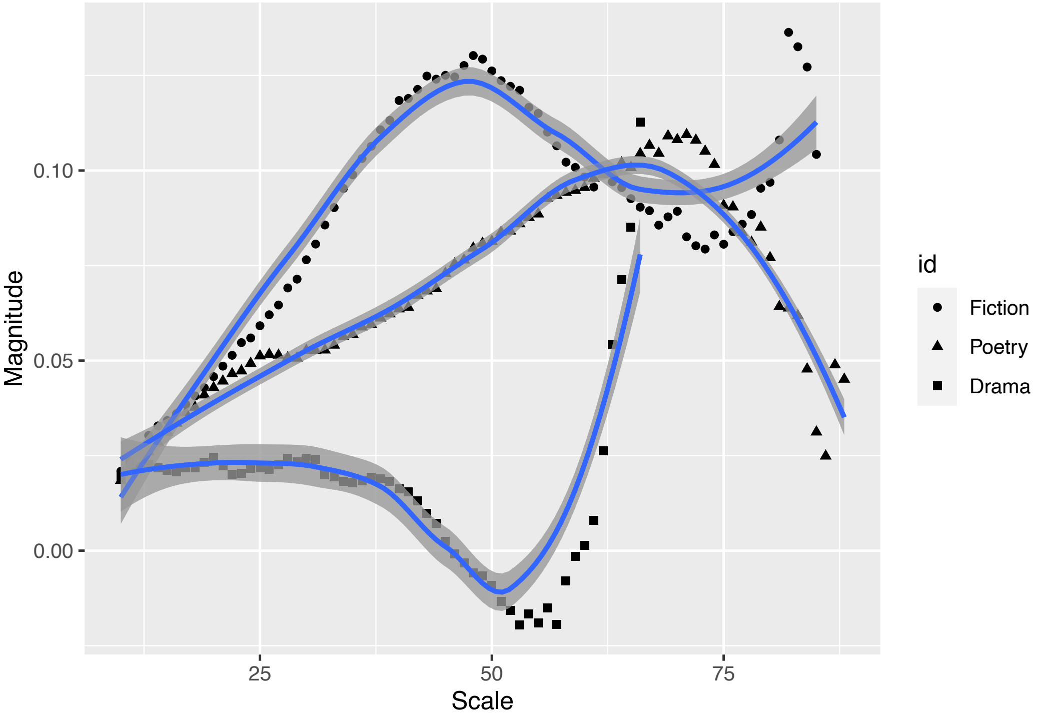


Figure 2.7: Magnitude of significant t-tests for period of time

Before discussing these plots in detail, it is important to note that the line and confidence bands superimposed on the data points represents the results obtained from a simple regression of y ~ x, implemented through one of ggplot's 'loess' smoothers for the purposes of rendering the visualisations more coherent and in order to provide some indication of the data points' direction of travel.

In many respects, the results outlined in *Figs.* 2.6 and 2.7 conform to logical expectations. Though the trend lines, which we see passing through each of our data points are often erratic, in very broad terms, the broader the scale at which we analyse distances, the more likely it is that i) we will detect significant differences and ii) these differences will be more pronounced. Poetry, represented by the triangles and the trend line which passes through them, seems to be the one genre in which this is not the case, both percentage and magnitude of significant differences decline significantly after a scale of 65 years. This leaves us with the question as to how exactly to identify a 'break' in order to proceed with our analysis, as both of our *Figs* give us significant room for interpretation. We could choose to analyse scales which are capable of returning significant differences 100% of the time, but this might forestall the opportunity of identifying breaks which take place over shorter periods of time. In pursuit of a benchmark, we borrowed an approach from machine learning and identified the point at which each of the trend lines in *Fig.* 2.6 exhibited an 'elbow' in its curve to avoid overfitting the model of temporal change as it were. These three elbows in each trend line are identified in *Fig* 2.6 by the vertical lines drawn from the bottom of the graph to the trend lines themselves. By doing so, we identify a break in the fiction corpus at 57 years, 50 years in poetry and 38 years in drama. These are illustrated by the vertical lines in *Fig* 2.6. As can be seen in Table 2.15, these were the estimates which were ultimately opted for.

In addition to potentially allow us to identify the breaks, these scales have the advantage of being roughly commensurate with the turnover of established literary-critical hypotheses. In very broad and perhaps even simplistic terms, we might say the period of 1751 to 1922 covers roughly three literary epochs, from neo-classicism from the mid to late eighteenth century, realism and romanticism in the early to mid nineteenth century and modernism of the mid nineteenth century to the early twentieth. We would therefore expect in and around 3 distinctive epochal transformations across the chronology of these datasets, which seemed plausible and to broadly cohere with our quantitative findings so far.

This chapter began with a consideration of the corpus which would be best suited in order to answer the question as to what rate literary change can be said to occur. It provided an extensive rationale for why the corpus which was eventually identified, a database of word frequencies obtained from the HathiTrust was best suited to this goal, by virtue of its size and time span. We then considered the contents of the dataset itself and how a certain degree of cleaning and other forms of data manipulation were necessary in order to render it more suitable for an analysis. The distances existing between each year, calculated on the basis of relative frequencies of MFWs, were analysed via the application and visualisation of results arising from a series of *t*-tests. Having established these benchmarks for each of the three genres existing in our datasets, the next chapter will analyse the results themselves, through use of our benchmarks and identify the words which may be said to be shaping or determining the nature of these breaks themselves.

Chapter 3

Results

This chapter will detail the outputs of this thesis' quantitative approach and can be divided into three parts. First, we perform a series of correlation tests on Delta distances both forwards and backwards in time in order to examine the temporal dynamics associated with the emergence and transmission of literary novelty over time. As in the previous chapter we identified a series of benchmarks within which we would expect to encounter quantitative evidence of a distinct change of direction in the literary history record, a break, we can now take these benchmarks into account when analysing the data itself. The third and final part will describe how regularised logistic regression was used in order to identify words symptomatic of literary production both before and after a break can be identified as having taken place.

3.1 Correlations

In an approach derived from Barron et al., we identified distance backwards as novelty, the amount of difference a particular year introduces, and distance forwards as transience, the amount of difference which subsequent years introduce.

A simplified version of how this works in practice appears in *Figs.* 3.1 and 3.2. In *Fig.* 3.1 we can see that 1751 introduces a distance of 2 from 1750 and that 1752 introduces a distance of 8 from 1751. As these distance measurements are distance backwards in time, we refer to them as novelty.

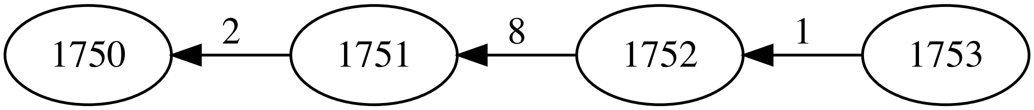


Figure 3.1: An diagram illustrating novelty over a four year period

In *Fig.* 3.2 we see that 1751 sits at a distance of 7 from 1752 and that 1752 exists at a distance of 6 from 1753. As these are distances forward in time, we refer to them as transience.

One of Barron et al.'s approaches for identifying influential agents within a corpus is to subtract each data point's transience from its novelty in order to attain a third statistic, resonance. By subtracting 1751's transience from its novelty, we find that 1751 has a resonance of -5 (2 - 7 = -5) and by subtracting 1752's transience from its novelty, we find 1752 has a resonance of 2 (8 - 6 = 2). Therefore, we can say that 1751 scores low for resonance, whereas 1752 could potentially score quite highly.

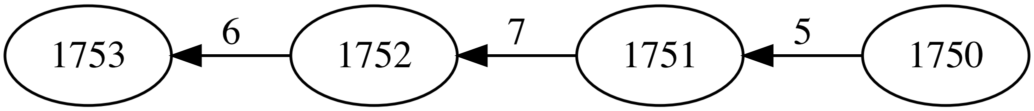


Figure 3.2: An diagram illustrating transience over a four year period

Of course this example does not strictly make logical sense. Firstly we are attributing two different numbers to the same distance and secondly the distances we are analysing are several decades in size, but it will be sufficient to illustrate the principle that if a particular year is more novel than it is transient, it will be more closely aligned with subsequent than with previous years, rendering it influential. If it is more transient than it is novel, we will surmise that the differences it has introduced will have failed to sustain themselves.

Table 3.1: Novelty Correlated with Resonance (Fiction)

	Date	Novelty	Transience	Resonance
Date	1.00	0.21	-0.97	0.89
Novelty	0.21	1.00	-0.29	0.59
Transience	-0.97	-0.29	1.00	-0.94
Resonance	0.89	0.59	-0.94	1.00

Table 3.2: Novelty Correlated with Resonance (Poetry)

	Date	Novelty	Transience	Resonance
Date	1.00	0.69	-0.86	0.86
Novelty	0.69	1.00	-0.69	0.87
Transience	-0.86	-0.69	1.00	-0.96
Resonance	0.86	0.87	-0.96	1.00

In each dataset, novelty, transience and resonance were calculated for each year according to the relevant statistical baseline as can be seen in Table 2.15. These three variables were normalised by the mean in line with Barron et al.'s approach. The relationships which exist between these observations as well as the passing of time was then investigated via a series of correlation tests.

The results of these correlation tests can be seen in Tables 3.1, 3.2 and 3.3. Each of these tables are laid out in the form of a matrix so that the correlation co-efficients between the four variables we are concerned with analysing, date, novelty, transience and resonance, can be easily identified. There is a broad spectrum of results to be observed here, but it is nevertheless possible to draw some preliminary conclusions on the basis of the behaviour

Table 3.3: Novelty Correlated with Resonance (Drama)

	Date	Novelty	Transience	Resonance
Date	1.00	0.30	0.00	0.17
Novelty	0.30	1.00	-0.50	0.85
Transience	0.00	-0.50	1.00	-0.88
Resonance	0.17	0.85	-0.88	1.00

of the correlation coefficients.

The first observation we might make is that the fiction and poetry datasets are broadly similar in terms of the observed correlation coefficients which are statistically significant and here we associate statistical significance as is customary, with correlation co-efficients higher than or equal to 0.7 or less than or equal to -0.7. We see in Tables 3.1, 3.2 that in the negative correlation between date and transience (-0.97 in the fiction dataset, -0.86 in the poetry dataset) and the positive correlation between date and resonance (0.89 in the fiction dataset, 0.86 in the poetry dataset) that as time passes, the dataset becomes less transient and more resonant. We might then propose that what we are witnessing here is a certain degree of convergence of prose and poetic form over time. The one difference existing between poetry and fiction, is the way in which resonance appears to be highly positively correlated with novelty. It may then be said that novelty in poetry is rewarded with more resonance than it is in the fiction dataset. Though the correlation co-efficient between novelty and resonance fall short of statistical significance as we interpret it here, at 0.59 it is not a totally negligible effect size. We can also see this greater degree of reward for novelty in the drama dataset, where novelty is also positively correlated with resonance.

While in the early stages of this project it was hoped that a broad-scale analysis of maco-economic dynamics such as those that we see here could easily lead onto more qualitative readings and the identification of the most influential agents within the dataset in highly specific terms. Just as Barron et al. identify particular speakers in the French revolutionary debates as setting the agenda, it was hoped that we could hone in on individual works by individual writers and find a means of proving through empirical analysis that what our database provides an insight into is the stabilisation of these three forms and the greater capacity for poetry and drama to reward innovative interventions into the literary-historical record, with the additional finding that poetry and drama are significantly less indebted to norms of usage as one would observe within prose, especially by the onset of the twentieth century. However, once the attempt was made to hone in on particular works through examinations of the metadata at crucial points significant difficulties began to arise. As has already been mentioned, much of the word frequency data held by the HathiTrust arise from materials contained within a wide array of digital libraries and in many instances works which form a crucial part of our analysis over entire decades are simply not available. Even in instances where they are, they were created through an automated OCR pipeline and almost totally illegible. However, without further research or work undertaken in this area, this hypothesis may remain standing as a largely speculative account of these results, open to any future scholar to investigate further.

3.2 Identifying breaks

As this thesis is invested in identifying breaks, we were most interested in identifying years which score highly for both novelty and resonance, where scoring highly is understood as two or more standard deviations above or below the mean, in line with Barron et al.'s approach. We would therefore be most interested in instances which are highly novel and highly resonant. Of course, as Barron et al. demonstrate, the existence of this combination implies the existence of three other combinations which might be similarly illustrative. We might therefore also investigate instances in which i) high novelty and low resonance is observed, ii) low novelty, high resonance is observed and iii) low novelty, low resonance is observed. Though these three categories are inherited from Barron et al.'s approach to categorising individual speakers within the French revolutionary debates, they nevertheless present some interesting potential as case studies for the transmission of literary influence.

If we identify high novelty and high resonance as indicative of 'breaks,' years which introduce significant differences and are correspondingly influential, high novelty low resonance might be regarded as an unsuccessful attempt to break with the past, a period in which some new words or styles began to come into prominence, but fail to survive

Table 3.4: Proposed different types of breaks

Novelty	Resonance	Type	Proposed Explanation
< 2	< 2	Break	Introduces significant differences, correspondingly influential
< 2	> 2	Failed Break	Unsuccessful attempt to break with the past
> 2	< 2	Tradition	Persistence of tradition over time
> 2	> 2	Null	No significant difference from predecessors, also not influential

to the same extent. A year in which a low amount of novelty is observed but there is nevertheless high amounts of resonance, might be said to represent the persistence of tradition over time. Finally, we might propose the existence of a year in which there is both low novelty and low resonance, one which is homogenous to the degree that it fails to mark a significant difference from its predecessors but is also not influential, presumably because it is overtaken by a more resonant year very soon after or before. These four potential combinations of values are illustrated in a more straightforward way in Table 3.4.

In order to identify some of these variables or combinations of variables, we extracted every instance of novelty and resonance which scored higher than 2 or lower than -2 from our data, as can be seen in Tables 3.5, 3.6 and 3.7. In line with Barron et al.'s methodology, we also applied the Kullback-Leibler divergence to each year in order to arrive at an alternate means of calculating novelty, trasience and resonance in such a way which might form a contrast with Delta. However, as z-scored Kullback-Leibler divergence was unable to identify any significant divergences from the mean from the perspective of novelty or resonance, it was decided that we would confine ourselves to a consideration of the results obtained via Delta.

As it turned out, the most consistently occurring of our four envisioned types of year is the category of low novelty low resonance, which occurs in the fiction dataset in 1820 and in the poetry dataset in 1810. The break of high novelty and high resonance within which we are most interested, does not occur. Therefore, none of the transformative changes which we envisioned before this analysis was initiated can be successfully identified.

Table 3.5: Breaks (Fiction)

Date	Novelty	Transience	Resonance
1820	-2.78	1.58	-2.30
1823	2.09	1.04	-0.15
1825	2.17	0.45	0.38
1826	-2.32	0.88	-1.54
1835	2.34	-0.34	1.10

Table 3.6: Breaks (Poetry)

Date	Novelty	Transience	Resonance
1810	-2.87	1.46	-2.16
1813	-2.35	1.09	-1.70
1825	-2.50	1.04	-1.72
1864	1.33	-2.50	2.22

Table 3.7: Breaks (Drama)

Date	Novelty	Transience	Resonance
1848	-2.35	0.00	-1.31
1858	1.71	-2.04	2.17

Failing to identify the presence of proper breaks in the datasets was obviously a disappointing result, which in certain respects forecloses any attempts we might make to claim the findings contained within this thesis are in and of themselves, novel. Nevertheless, it was decided to take the years which most closely approximate a break, those which score highest for novelty and resonance for further study. The closest approximation of a break therefore appears in the fiction dataset in 1835, 1864 in poetry and 1858 in drama. All of these take place within a twenty-nine year time frame, in and around the dataset's halfway point.

3.3 Regularised regression

The next objective was to identify which textual changes these years can be said to introduce. The best means of answering this question is to identify which words are introduced or are deployed more often after the break than before, based on our understanding of approximately how long it takes for a transformative change to be noted within each of our datasets, even if these changes are not quite as transformative as we initially projected. As we are interested in the emergence, increased or decreased usage of words in one period of time as opposed to another, the next objective was to render the information obtained in the tables above relevant to particular word frequencies themselves. There are a number of ways in which this could be accomplished. We could make use of *t*-tests in order to identify whether or not the mean difference from the perspective of word usage is significantly different in one period of time as opposed to another. Another would be to reduce the number of variables, as is done in PCA, and combine our word frequency vectors into broader components. However, as we wish to retain our variables as individual words and we want to avoid repeating the method used earlier, we might be more interested in identifying words which can serve as predictor variables, a word which, given its pattern of usage, we could use in order to make an accurate estimation as to whether the year is more likely to take place before or after a 'break' year.

The aim of regression in general is to model the relationship which exists between one dependent and a number of independent variables. In this instance, our independent variables are word frequencies and our dependent variable is a categorical one; which of the two phases of time is this year most likely to belong to. We assume that the answer to this question is contingent on the relative frequencies of the words and we therefore aim to construct a mathematical function which can best fit the variation of the data (Moisl 135–36). Logistic regression is then the most suitable regression method to adopt in this instance as we are dealing with a binary classification task (Lantz 191): on the basis of the cross-section of words which appear in this year is it more likely to belong within a pre as opposed to a post-break milieu?

It is axiomatic within ordinary least squares regression (OLS) that models should be parsimonious and that as few variables should be regressed along the dependent variable as possible (Crawley 8). The aim of regularised regression remains the same as OLS regression, the minimising of error is the overriding objective. This is accomplished by finding a penalty term, referred to as lambda, by which one reduces the value of particular variables (McDonald 99). However, since we have thousands of independent variables, many of which are collinear with one another, minimising our SSR is not a definitive solution (Hoerl and Kennard 55). It would be straightforward to regress for five or so variables, but thousands and thousands will very quickly create a saturated model, which would be 'overfit' and would not perform well upon the test data, as it will have been too rigidly trained in differentiating one block of training data from another. One solution to the problem of collinearity might be to remove particular variables and effectively delete words from our model (McDonald 93). However we do not wish to exclude words which might be relevant to the model before we have even begun. In order to carry out a regression with all of these variables within the model, we therefore need to adopt a regularisation method, which is suited to analysing large numbers of collinear variables and does so by constraining their magnitude, applying penalties to the predictors, introducing bias and reducing variance (Çiftsüren and Akkol; Lever et al.). All of the most commonly used forms of regularisation, such as elastic net, ridge regression, least absolute shrinkage and selection operator regression (LASSO) are based on the application of penalties to particular variables and thereby enhancing the capacity of the model to make predictions. These techniques are most often used in order to analyse high-dimensional data that occur most commonly within biostatistics or genomics (Ogutu et al.; Waldron et al.), where the object of phenotype or breeding pair prediction is complicated by a potentially vast number of independent variables, many of which are either highly correlated with one another, irrelevant to the dependent variable or far exceed the number of samples in quantity. The appeal of regularisation techniques is that under circumstances, which map well onto the profile of word frequency data, they can facilitate the creation of models which can generalise very effectively to new data. Ridge regression is one such approach which functions by shrinking the co-efficients of correlated predictors towards zero, but does not remove them from the model altogether (Ogutu et al. 2). LASSO does reduce particular variables to zero and LASSO therefore functions more as a means of variable selection. Due to the fact that LASSO can tend to choose a single group of collinear co-efficients and ignore the rest of the variables in quite arbitrary ways, LASSO is not as robust against high correlations as ridge regression is (Waldron et al. 3399). Hui Zou and Trevor Hastie therefore developed elastic net, which aims to combine the best of both ridge and LASSO regularisation within one algorithm. Elastic net combines a ridge regression into the penalty term in order to make the LASSO shrinkage more robust against the pitfalls which are usually associated with LASSO. In summary, elastic net and LASSO are two approaches which are alike in that they remove variables from the model by penalising them to zero, a significant aspect of their approach involves their functioning simultaneously as a form of parameter selection (Çiftsüren and Akkol 280; Lever et al. 804; Ogutu et al. 3). As we wished to retain all our variables within the model and leave ourselves with as many word-types potentially associated with the breaks as possible, we opted for ridge regression and this is the regularisation technique which was ultimately adopted.

Table 3.8: Regularised regression results

Break	Genre	Sample	Accuracy
1835	Fiction	20%	97.2%
1864	Poetry	20%	95%
1858	Drama	20%	80.9%

To give a tangible example of how this method works in practice, we begin by choosing a particular break year from Table 3.5, in this instance, 1835. As we have already determined that we expect a break in the fiction dataset over margins of fifty-six years, we partition our data into two separate blocks of equal size. The first block contains the fifty-six years before 1835, every year from 1779 to 1834 inclusive. The second block contains every year from 1836 to 1891 inclusive. It is important to note that the break year itself is not included within either of these two blocks, as we are interested in the effects that might plausibly be attributed to its influence as opposed to the discrete contents of that particular year itself. After some experimentation with different sampling rates, we found that 20% minimises the likelihood that a particular year from our pre-break data would be mistaken for a post-break year and vice versa. Therefore, we randomly sampled 20% of both the first and second block. We then perform a cross-validated fit, which tested lambda across a range of values in order to identify the penalty term which reduces the SSR to the greatest extent. This process was executed within a loop run one hundred times in order to calculate a representative confusion rate. The words that acted as predictors, as well as the number of times that they did so, were also obtained and the results from these analyses appear in Table 3.8.

Our use of regularisation is complicated slightly by the fact that we are not attempting to hone in on a single or straightforward answer, we are not interested in identifying one word which outperforms all others. Instead, we are attempting to identify a set of words which correlate with the relative certainty that a particular year belongs to one block of years in one timeframe as opposed to a second block in another. In broad terms we therefore take these results which appear in Table 3.8 as indicative of the fact that

Table 3.9: Predictor variables

Year	Genre	Positively Correlating Words	Negatively Correlating Words
1835	Fiction	618	746
1864	Poetry	291	118
1858	Drama	144	35

it is possible to separate two groups of years on the basis of the relative frequencies of words used. We also note that it is in the drama dataset, where we did not identify anything even closely resembling a break, that we get the lowest results for accuracy, by approximately 15%.

3.4 Quantifying words

It is axiomatic within CLS that mediating between macro and micro scales of reading presents extensive difficulties to the analyst as it requires oscillating between micro phenomena and the broader macro theory within which we are attempting to render them coherent. This thesis was in no way immune to this problem, but contends that of all potential solutions, the one it ultimately came to was largely satisfactory in this respect. Overall details relating to our predictor variables which were obtained from the results of our regularised logistic regression can be seen in Table 3.9.

We can propose at least three findings regarding the relationship between breaks and predictor variables across genre based on the results displayed in Table 3.9. The first is that in instances where there is not a ‘break’ present — in the sense of both high novelty and high resonance being present in a particular year at once — significant numbers of predictors may still be identified. The second observation or finding is that two out of the three breaks introduce more words than they take away; the number of positively correlating words in poetry and drama exceed the number of negatively correlating words by more than two times. The exception is the fiction dataset, in which the number of negatively correlating words exceeds the number of positively correlating words by approximately 130. In straightforward terms then, we might say that the introduction of new words does not correspond to a reduction of words on a similar scale, a break is rather an event which facilitates greater amounts of heterogeneity.

The results that we have obtained here as regards our predictors would seem to contradict the results of the correlation tests obtained in section 3.1, where our results suggested a solely negative sense of historical change, where a stabilisation or perhaps even a rigidity of form begins to take shape towards the end of the nineteenth century. Here, in the number of predictors being overwhelmingly positively correlating as opposed to negatively points to an oppositional, positive dynamic which suggests periods of literary production are more characterised by growth and expansion rather than specialisation or the adoption or consolidation of formally constraining rules.

In attempting to move these sequences of results towards a conclusion, though our methods seems to be capable of identifying particular words which can be regarded as correlating to changes in literary production over time, it would be difficult to maintain the sense of a break which was defined in this study's opening sections. It seems far more accurate to characterise these breaks as symptoms or moments within a broader objective tendency which operates far more steadily over a far longer timeframe. Breaks might correspond to the trend, but they are not independent givens, and literary change does not seem to take place in a manner which is conjunctural or fluctuating. This longer pattern becomes particularly clear when we graph the percentage of texts produced in each year that contain words which correlate with breaks, both positively and negatively. The graphs appear in *Figs.* 3.3, 3.4 and 3.5.

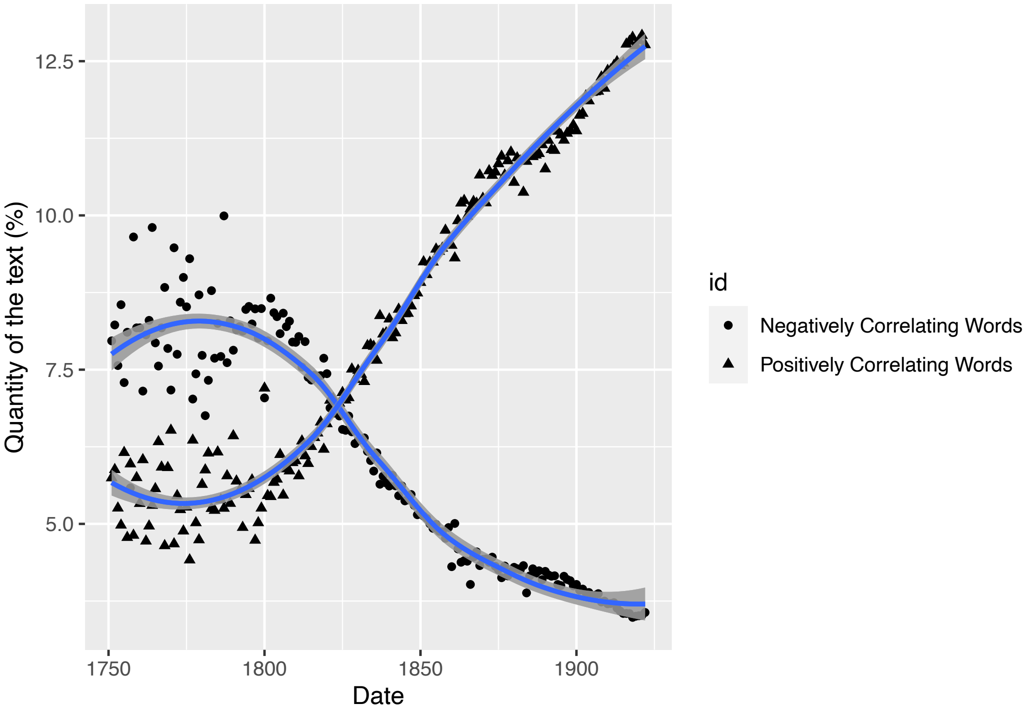


Figure 3.3: Percentage of the corpus composed of break words (Fiction)

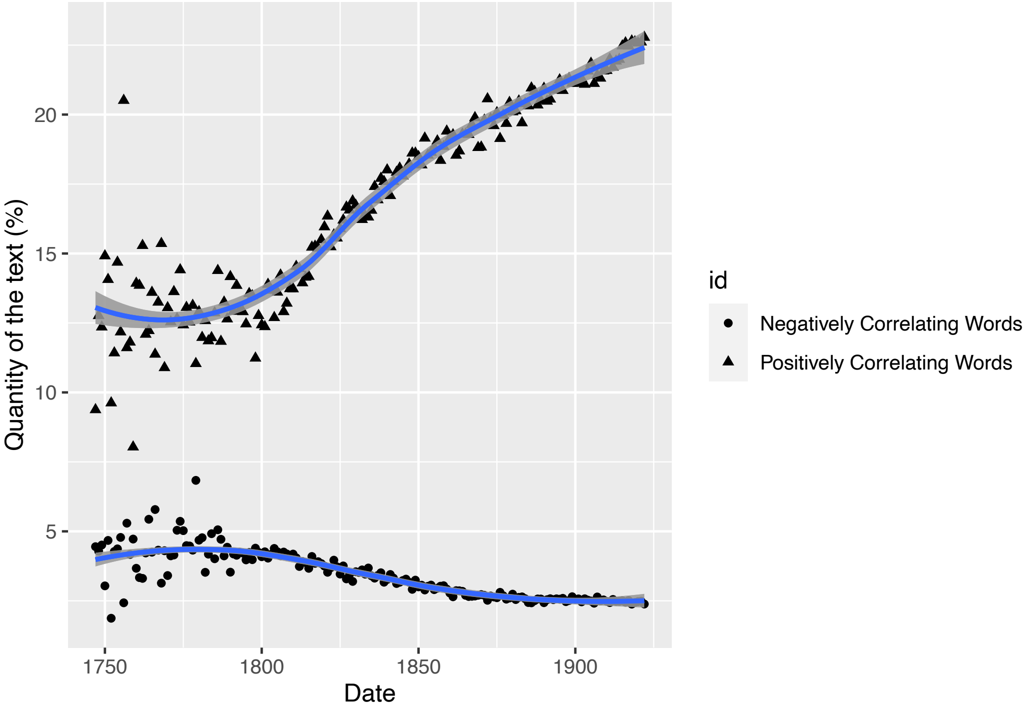


Figure 3.4: Percentage of the corpus composed of break words (Poetry)

In *Fig.* 3.3, we see 618 words increase their overall incidence in the fiction dataset by approximately 7.5% over a 171 year period. In *Fig.* 3.4 we see a smaller cohort of words, 291, increase their overall incidence in poetry by 10% over the same period of time. In *Fig.* 3.5, 144 words double their incidence in drama from ~6% to 12.5% between 1791 and 1922. The differences between the fiction dataset on the one hand and the poetry and drama datasets on the other are clear. While in drama and poetry the two lines representing the quantity of the text commanded by either positively or negatively correlating words are significantly differentiated, indicating a very clear bifurcation of two reasonably distinct

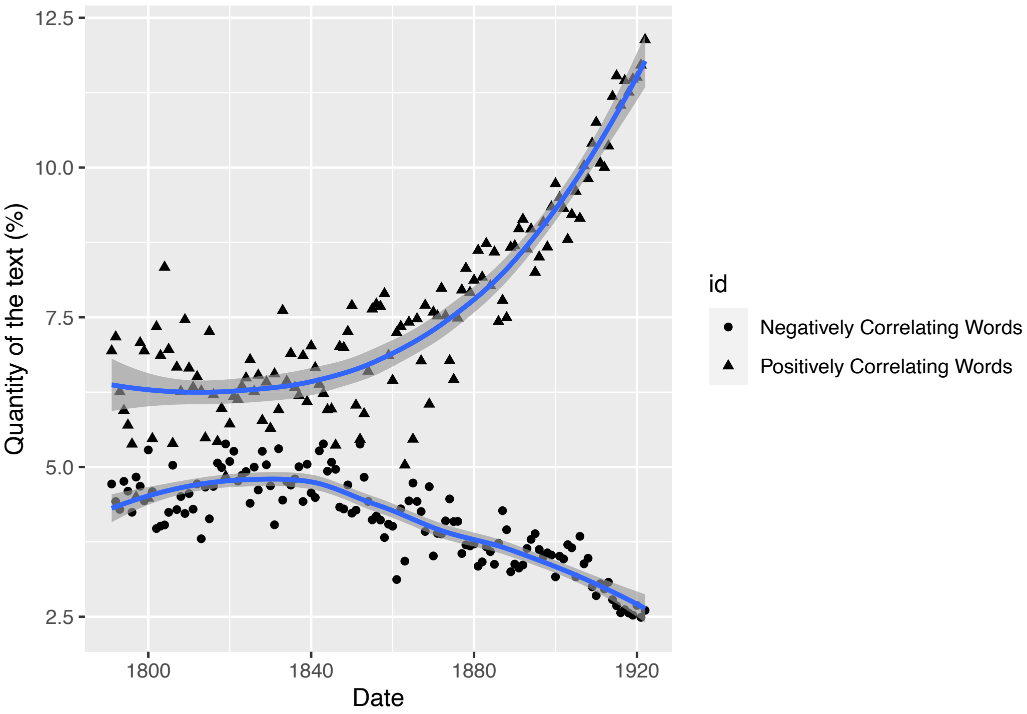


Figure 3.5: Percentage of the corpus composed of break words (Drama)

semantic fields, we see a far higher amounts of intermingling in the fiction dataset. This takes place to the extent that the trend lines cross one another in and around 1825, which seems to mark the point at which one begins to assert its difference from or supercession over, the other. What we seem to see here is the triumph of one formal modality in fiction over another. In poetry and drama, we see the word use associated with one tradition decline while word use associated with another tradition grows.

The determining factor lying behind this difference will be taken up in the next chapter, but the most important point or finding to underline at this stage is that the Underwood hypothesis of incremental change remains robust. Crucially, it does so even when methods which are specifically directed towards the identification of variables associated with periods of accelerated change are used. In short: words do not stop being used suddenly and are not introduced into the dataset suddenly, but do so incrementally.

3.5 Interpreting words

Once we had obtained our predictor variables as documented above, it was then necessary to devise a means of reading them coherently within a literary-critical hypothesis of historical change. This was a more complex undertaking than it might seem at first. Before discussing the results obtained in and of themselves, we will provide an account of each adopted mediating approach and why each one was ultimately found to be unsatisfactory in one or another key respect.

At first we experimented with POS tagging, a method used in Natural Language Processing (NLP) and linguistic analysis in order to characterise patterns of word usage from a grammatical perspective. While many POS taggers are distinct from one another and draw on different vocabularies or incorporate machine learning to varying extents, the Penn Treebank POS tagger would be among the most well-known and most widely used of these due to its incorporation in the Natural Language Toolkit (NLTK); a collection of libraries and programs used for text analysis in the Python programming language

(Marcus et al.). The Penn Treebank allows us to sort words into one of thirty-six different grammatical categories such as predeterminer, singular proper noun, or comparative adjectives. Given that POS tagging allows us to introduce another layer of abstraction which might be illustrative on a broad scale, but furthermore legible from a literary-critical perspective, it was anticipated that POS tagging would be a productive approach to adopt. However as it turned out, there was relatively little to be gained from analysing words from the perspective of their patterns of grammatical usage from one epoch to another. The differences in amounts of verbs, adjectives or nouns were found to be marginal and furthermore difficult to interpret qualitatively. It seemed then that it was not another layer of abstraction which was required at this stage of the analysis but instead an approach which would allow us to keep the words themselves and their content most clearly in view.

We then experimented with a wholly qualitative approach, which involved reading the words individually from the perspective of their content alone, giving particular attention to words which appeared as predictors. This approach presented its own difficulties, reading the words in this way would be a largely subjective endeavour. In addition, words are inherently polysemous and their significance can vary enormously depending on context, a context which is unavailable to us due to the way in which the HathiTrust collection has been compiled. If we consider the word 'back' for example, which increases significantly in the fifty-one year period after 1858 within the drama corpus, it is unclear whether this word's increased usage refers to the body part, the far end of something, or a verb. It was obvious that we needed to take steps in order to re-create some of the absent context from our dataset. In attempting to do, we turned to the CLS paper whose approach most closely approximates this project's approach, namely, Ryan Heuser and Long Le-Khac's analysis of nineteenth century British novels, produced within the context of the Stanford Lit Lab. Heuser and Le-Khac had significant amounts of success in constructing qualitatively illustrative perspectives on literary history by correlating word usage and constructing broader 'semantic fields' on the basis of significant correlation coefficients (Heuser and Le-Khac). It was thought that by using word correlations in the same way, we could move beyond the dangers associated with reading individual words from the perspective of their content, overriding individual words with multiple meanings by honing in on the broader patterns which might lie behind their usage, a machine assisted middle ground. Perhaps unsurprisingly, words which correlated positively with 'back' in the drama dataset were already returned in our positive correlation results and words which correlated negatively had already emerged in our negatively correlating words. In instances where statistically significant correlations were detected, more than half of these overlapped to a significant extent with the word-types identified through regularised regression. Furthermore, these word-types in many ways left us where we had started; with significant numbers of words and no means of accounting for them from a more qualitative point of view.

It was then decided that the best means of reading this data in such a way that would allow us to formulate a robust and distinctive literary-critical hypotheses would be to read the data back into their original context by locating full digital versions of texts from key points within the ~170 year period we are analysing. In the next chapter, we will see how sections of these texts which are particularly dense with words which are either positively or negatively correlating were analysed via computational means.

3.6 Conclusion

The first finding that this chapter came to through a series of correlation tests is that the fiction and poetry datasets are both alike in that as time passes, both datasets become less transient and more resonant, perhaps indicating that there is a convergence, or a continuous building towards a formal norm. In poetry however, we see resonance is highly positively correlated with novelty and therefore it is possible to speculate that novelty is rewarded in poetry to a greater extent than it is in the fiction dataset. As regards findings of more direct relevance to this thesis' primary research question, having identified for each of our three literary genres specific intervals in which we would expect to see a transformation of the literary milieu, it is possible to say that there are no significant breaks to be observed in literary history in the sense intended at the outset of this thesis. Rather, the words denotative of a changing milieu over time, while they point towards an expansiveness and increasing incorporation of a broad vocabulary, do not in any sense arise suddenly. We then documented the issues which arose in our attempts to undertake readings which were qualitative in nature, such as redundancy in the case of further correlation tests, polysemy in the case of words alone and terseness in the case of POS tags. The next chapter will seek to provide a series of qualitative-based readings having taken the shortcomings of these prior approaches into account as well as examine the consequences of our results for literary criticism in general.

Chapter 4

Prospects

4.1 ‘Breaks’ and literary change

So far, this thesis has outlined the capacity of John Burrows’ Delta method to identify periods of accelerated literary change when applied across time rather than across text. To summarise the method as it was applied in the second chapter of this thesis, we first represented each year as a vector of the relative frequencies of the third quartile of most frequent words produced in a given year and then calculated the Delta distance between these years. We then applied *t*-tests to the resultant distances in order to identify a benchmark within which we would expect to see a qualitative transformation of the literary history. In an approach derived from Barron et al., we identified distance backwards in time as novelty, the amount of difference a particular year introduces, and distance forwards as transience, the amount of difference which subsequent years introduce. By subtracting each data point’s transience from its novelty, we obtained a third statistic, resonance.

As it turned out, years in which high novelty and high resonance which we postulated as existing at the start of this study, in which we are most interested, do not occur.

Therefore, none of the transformative changes which we envisioned before this analysis was initiated can be identified. Nevertheless, it was decided to take the years which most closely approximate a break, those which score highest for novelty and resonance for further study. The closest approximation of a break therefore appears in the fiction dataset in 1835, 1864 in poetry and 1858 in drama. Through the application of regularised regression, we trained a model to differentiate years which more accurately fit the profile of a pre or post-break milieu. We then obtained a figure between zero and one which scores the degree to which each year fits one or the other category and we correlated this statistic with our word frequencies in order to identify which words are indicative of these transformations. By quantifying and visualising the extent to which the texts produced in a given year consist of these either positively or negatively correlating words, we can see that literature does not change in the sudden or conjunctural manner which we envisioned at the project's outset. As can be seen in *Figs.* 3.3, 3.4 and 3.5 in section 3.4, the words which are associated with these points in time increase and decrease incrementally rather than quickly.

The Underwood hypothesis of incremental literary change which this thesis initially set itself the task of testing has therefore been validated. Not only has Underwood been proven correct in his demonstration that the rate of literary change is slow, but if we look to Appendix E, we can see that the findings Underwood also presents regarding the content of modern literary production were also replicated. We can observe an increase in words associated with dialogue ('said,' 'asked,' 'say') and sensory perception ('touched,' 'see,' 'listen') along with an attendant decrease in words associated with politics ('liberty,' 'reason,' 'title'), economics ('lodged' 'owed,' 'procured'), and religion ('blessings,' 'heaven,' 'graces'); two trends which broadly align with the rise of the genre known as literary realism. Underwood furthermore demonstrates that these trends are not only capable of encapsulating literary production from 1750 to 1922, but that they seem to continue right up to, and presumably persist beyond, the year 2000 (Underwood, *Distant Horizons: Digital Evidence and Literary Change* 23–25).

Having accounted for these changes from an empirical point of view, this chapter will examine the finding that the past three hundred years of literary production are best rendered as the continued rise of realism in opposition to the work of Raymond Williams, Terry Eagleton and Frederic Jameson. The most influential critical accounts of literary history produced by these critics have placed disproportionate emphasis on breaks, especially one which is supposed to have taken place between nineteenth-century realism and twentieth-century modernism. In doing so, they have made extensive use of political, economic and historical works produced by Karl Marx as well as other works produced within the broader Marxian tradition, due to the robust conceptual framework it offers for understanding industrialisation and its social and political ramifications. However, by virtue of the fact that their foremost argument, that twentieth century literature is significantly different from literature produced in the nineteenth century does not seem to have been borne out by the historical evidence in both this study and Underwood's, it is necessary to return to them and consider what aspect of the Marxian framework can be retained or salvaged in our account of modern literary production.

After definining exactly what is meant by realism and modernism, this chapter will locate the origin point of this over-emphasis of the difference between the two moments in literary history in the ideological predispositions of the first institutionalised schools of literary criticism of the early to mid twentieth century. Marxian critics such as Williams, Eagleton, Jameson and others working from the sixties onwards within the context of cultural studies and the New Left sought to overcome such ideological readings and to integrate more overtly political concerns into their literary critical practices. Their failure to overcome the ideological formulation of the twentieth century break can be attributed to the abiding influence of the philosophy of George Wilhelm Friedrich Hegel in Marxian literary criticism. Hegel's philosophical system exerted a significant influence on Marx's own comprehension of the social totality and the role of capitalism as a mode of production within it (Liedman 33). It may be argued that such a reversion was inevitable for any criticism which attempted to take Marx's writings as foundational, given the relative lack of actual works written by Marx, or his lifelong collaborator Friedrich Engels, which consider culture or processes of cultural production. Critics seeking a means of integrating Marxist theories into their critical methodologies were therefore by necessity thrown back onto Hegel's system with its highly involved and intricate means of describing processes such as mediation, reflection and alterity in a manner which lends itself to readings of the text as developing rationally out of its own concepts in the process of being read or interpreted. What this reversion overlooks is Marx's break from Hegel's philosophical system. Central to Marx's particular social critique as well as the radical intellectual circles in which Marx formulated it, was the rejection of Hegel's foundational assumption that Spirit represents the primary agent of world history, emphasising collective social formations arising from within actual human societies to a much greater extent (Liedman 89).

The central objective of Marxian literary criticism, to account for the ways in which historical progress is mediated in literary art, remains the most robust means of conducting an historical literary criticism, especially when empirical methods are involved, but it has yet to manifest Marx's break from Hegel. This has the result that literary form and content are read as more or less coeval with actual historical processes and our attention is therefore concentrated to a far greater extent on indeterminate processes of mediation rather than the historical processes which Marx identifies as determinative. The reification of textual rather than historical processes could be the reason why our empirical findings depart to such a great extent from the literary-critical historiography. This chapter will therefore look to materials Marx offers in the second and third volumes of *Capital* (1867) in order to explain the results we have identified and to formulate an explanatory model which affords the decisive role to historical and economic processes. Accomplishing this will furthermore involve the close reading of extracts from germane texts by literary authors such as Charles Baudelaire, William Wordsworth, D.H. Lawrence and Stephen Crane.

4.2 Realism and modernism in literary criticism

Realism is a mimetic mode of representation held by Ian Watt to have arisen in the late eighteenth century and, according to René Wellek's account, to have consolidated itself in the early nineteenth century. Some of its determinants include the development of empirical philosophy, a connection which is integral to Watt's account, as well as its content, representing as it does individuals within a clearly demarcated social context. Realism has therefore been promoted as literature's default mode insofar as an objective portrait of modern social reality is concerned. Indeed, Eysteinsson describes it as an effective zero-level against which subsequent aesthetic practices, associated with modernism, may be assessed. Literary modernism meanwhile, can be read as insisting on an opposition to realism's perceived aesthetic neutrality, and posits itself as a more formally self-conscious endeavour. This greater degree of aesthetic self-consciousness finds formal expression in the autonomy that form assumes within the works of many modernist authors and poets, such as James Joyce's *Ulysses* (1922), T.S. Eliot's 'The Waste Land' (1922) and Ezra Pound's *Cantos* (1922). Such works are contained or at least partially subsumed within a unified symbolic framework, derived from a classical as well as other mythological frameworks. We also see a certain degree of reflexivity in such works, an interest in representation conducted from an individual's point of view, and an overall integration of a stylistic restlessness as though the search for an adequate means of expression has become the formal logic of the work itself itself (Bradbury and McFarlane 26–29; Eysteinsson 186–92). While it will only form a significant part of our analysis later on in this chapter, it is important at this stage to identify literary romanticism as a key mediator between these two moments. Romanticism is yet another literary movement which first arises in the nineteenth century, in Germany and then England. According to M.H. Abrams, romanticism's emergence corresponds to a point in time in which neo-classical literary theory begins to give way to theories of expressive form, which emphasise the individual faculties and mind of the poet or author to a greater extent, paving the way for the sort of cult of the autonomous artwork operating in modernist literature (Abrams

20–21; Wilson 2).

While accepting that the first modern literary-critical schools were composed of quite distinct personnel spread over a variety of national territories, a shared tendency towards the valorisation of literary modernism, at the expense of its precursor realism, can be perceived as a more or less universal commitment. New criticism, practiced by F.R. Leavis and his followers was particular to Cambridge in the 1930's and 1940's, while a branch of American new criticism propagated by John Crowe Ransom, Cleanth Brooks, Allan Tate and Robert Penn Warren operated in the southern United States (Menand and Rainey 7). Formalism held sway in Russia, while structuralism emerged in France. Identifying a root cause for this shared valorisation of modernism necessitates a consideration of the imperatives guiding knowledge production in the Fordist state in the early to mid-twentieth century. This was an imperative within which the university played a key role, in establishing evidence-based methods for scholarship, while also providing an education to a population now attending universities in significantly higher numbers (Dickstein 323; Vinen 56). The new professionalism of academia, as well as the symptomatic literary-critical conception of the autonomous text, served the practical and ideological demands of an expanded university system during the cold war very well, especially in western states where administering the bureaucracy of the industrial state apparatus in a way which would match or exceed the capacities of the Soviet Union represented a political and economic imperative (Guy and Small 379; Vinen 63). These broader sociological changes are key in understanding how and why literary criticism moves from an appreciative and impressionistic form of appraisal more in the direction of a hard science, characterised by rigour, verification and the requirements associated with the weighing different standards of evidence (Day 167). Such methodologies were influential in elevating particular literary works with specific characteristics as opposed to others into prominence. Terence Hawkes for example, points to the significant influence Laurence Sterne's *The Life and Opinions of Tristram Shandy, Gentleman* (1759), with its use of verbal interplay and consistent tension between form and content, had in developing Victor Shklovsky's theory of defamiliarsation

(Hawkes 66). The essential commitment of these literary-critical schools of thought to the notion that literature and language operate as more or less autonomous and self-sufficient objects and that in considering them, historical, biographical and sociological fact, should play only a nominal, if any role at all, served to extend a preferential treatment to literary modernism as a project (Menand and Rainey 2; Frederic Jameson, *The Prison House of Language: A Critical Account of Structuralism and Russian Formalism* vi-ii). It is for these reasons that modern literary criticism begins to afford disproportionate amounts of attention to modernism within the history of literature.

John Brenkman, writing on the degree to which literary criticism has historically been implicated with its object of study, outlines the consequences of the blurring of the boundaries between literary criticism and the object of its attention. Brenkman argues that the rearguard of any given moment of cultural innovation, in this instance realism, comes by necessity to be read as socially conformist, naively empirical, or simplistically mimetic, in order to allow for a subsequent work or literary movement, to render human experience once again unfamiliar or estranged, to in effect, defamiliarise it (Brenkman 818). According to this argument, literary criticism becomes invested in reproducing narratives of progressivist evolution or supercession analogous to the ways in which modernism functions, a tendency which Williams has also identified (Williams, *The Politics of Modernism: Against the New Conformists* 3–7). Marxist literary critics working from the 1960s onwards aimed to challenge some of the assumptions undergirding the autonomy of cultural production in order to challenge the political conservatism of these approaches. Of course this did not take the form of a straightforward or open polemical conflict. Rather in specific national contexts, a significant amount of dialogue and interchange between the two paradigms represents the norm. In French literary criticism for example, we see Louis Althusser's capacity to fuse structuralist critiques of empiricism with Marxian science. In an English context, Eagleton has noted the continuities between Williams' studies of culture and Leavis' new critical emphasis on sensibility (Eagleton, "Criticism and Politics: The Work of Raymond Williams"). For the reason that Marx and Engels' collected works offer very little in the way of direct theories of culture or cultural expression, it was perhaps inevitable that the New Left would ultimately turn to Hegel's more abstract theories of mediation and historical development. In this way, the Hegelian dialectic became the primary means through which Marxian literary criticism operated. Proving this, before demonstrating how a new more Marxian departure is possible within the context of CLS will require a brief diversion on the nature of Hegel's philosophical system.

4.3 Hegel and literary criticism

At the basis of Hegel's philosophy is the development of consciousness. As Terry Pinkard notes, Hegel's primary object of study is therefore dual in nature; it is at once fixed, in that, broadly conceived it represents a single object of study that comes to be expressed and articulated in different ways over time. From this point of view, it is therefore mutable and dynamic as it proceeds along a potentially infinite series of inflection points (Pinkard 144). It is important to note that Hegel's philosophy is not limited to any narrow conception of human consciousness, such as that of a single individual or homogenous social collective coming into contact with any one thing outside of itself. Hegel prefers to regard consciousness as just one of the means through which more complex forms of thought are achieved, how the most basic and elementary forms of sense-data accumulate to form more complex assemblages, such as general categories within a system of logic, individual psychology, social and economic institutions, art, religion and even philosophy itself (Plant, "Hegel and Political Economy—II" 103). Michael Inwood provides an example of how this might work in practice, beginning with pre-historic man without the capacity for abstract thought, knowing only a sensuous existence, coming to develop the faculties for the consideration of concepts through the use of tools, or creative expression, as in cave painting. These objects therefore facilitate man's relation to his environment in more complex ways allowing him to mediate himself to himself, coming to self-consciousness and thereby superseding sensuous, unreflective and pre-historical being and potentially entering into more complex forms of thought, action and sociality (Inwood xv-i). Spirit is the totality within which these oppositions between internal or external, universal and particular, subject and object are reconciled as they develop subjectively out of their own concepts and ultimately become equal to their concrete or objective manifestations (Knox ix-x). To put this in more straightforward terms, it could not be said that an individual who desires freedom or agency within a particular social formation is free just because they desire it; it would be necessary for social development to be objectively adequate to the subjective impulse in order for an absolute freedom to become manifest.

Hegel's system represents a highly involved means of describing how Spirit and its predicates are deployed in developing self-consciousness in each of its stages. As, for example, the subjective desire for freedom becomes adequate to objective freedom and becomes absolute freedom, or how appearances in cave painting provide objects for reflection and eventually prove adequate to the development of the notion (Knox ix-x). For the purposes of our argument it will not be necessary to provide a thorough account of these processes as they documented in the *Phenomenology of Spirit* (1807) or *Science of Logic* (1812); it is obvious that this model of understanding in its actual component parts has played only a very marginal role in the history of literary critical praxis. In seeking to argue that it is a version of Hegel's philosophy as regards the state and the history of culture which is far more germane to the history of Marxian literary criticism than Marx's writings themselves, all that will be necessary is to underline at this juncture is that Hegel is a fundamentally teleological thinker and regards human history as tending towards the development of more harmonious and rational social arrangements. We see this when we compare Hegel's writings on the place of religion in Ancient Greek society with his writings on nineteenth century Germany. Hegel regarded the folk religion, as it was practiced in Ancient Athens, as providing the means through which the individual's desires could be mediated within the collective desires of the broader *polis* due to the specific ways in which religious ceremonies formed an integral and organic part of everyday social life, forming a stark contrast with religion as it functioned within Christian modernity, wherein highly cognised and intricate religious ceremonies have become removed from peoples' lived experiences (Plant, "Hegel and Political Economy—I" 80–81). Alienation from social practice is not a problem unique to modernity and it is furthermore integral to Spirit's development, but from Hegel's perspective, when industrial and commercial enterprises have developed within the straightened civic and religious institutions of the Holy Roman Empire, historically unprecedented contradictions have arisen. The solution to these contradictions is not to return to pre-Christian modes of existence or to roll back industrialisation however. The task of philosophy is rather to contribute to Spirit's advancement, to grasp the authentic nature of bourgeois Christian society so that these tensions may be reconciled within a new totality or synthesis, a task within which art, due to it being one of the ways through which man represents the world and Spirit comes to be embodied in sensory form, can play a crucial role (Hegel 8, 54–55; Plant, "Hegel and Political Economy—I" 85). As with Greek folk religion, Greek sculpture is Hegel's paradigmatic example of an art form which provided an apposite vehicle for man's stage of development at the time it did; a stage in human history in which there was no developed philosophy or theory of science. However, once specific paradigms of knowledge production arise in modernity, people are no longer capable of relating to art in the more primitive, sensuous form that the Greeks once did; art has become more cerebral and inwardly-directed. Hegel is here presenting a subtle critique of a regnant literary romanticism which he regarded as inadequate to the task by virtue of its interiority and focus on appearances; the work of advancing Spirit is more adequately played by philosophy (Hegel 81; Inwood xxv-ii; Knox vi).

In many respects the inclination of Marxian literary criticism in the direction of Hegelian philosophy was pre-determined by literary criticism conducted by critics working during the period of the Third International which tended to attribute disproportionate amounts of importance to the subjective experience of capitalist production, as though the effects that capitalism exerts on individual consciousness take primacy in any cultural account of them. Gillian Rose has written on the extent to which critics writing within the tradition of Western Marxism such as Walter Benjamin, György Lukács and Theodor W. Adorno misread Marx's conception of commodity fetishism arising within the capitalist mode of production, with the result that Marx's critique of the value-form and political economy in general are obscured. In the work of these critics and others writing in their wake, commodity fetishism and reification become mere shorthands for processes of objectification, rather than the social totality or mode of production within which these objectifications are taking place (Rose 38–39). Commodity fetishism is a concept which Marx introduces in the first volume of *Capital* in order to account for the ways in which capitalism's social relations are obscured by the outputs of the productive process. The capitalist mode of production forces the waged labourer to sell their labour power in order to produce a commodity, alienating them from their own productive powers at the same time that it immiserates them from an economic perspective. As can be seen in Appendix E, and as will be seen in the remainder of this chapter, the increasing presence of the industrially produced commodity in literature represents an integral part of our empirical findings ('hat,' 'chair,' 'cigar'). One instance in which real productive processes are elided in favour of a more experiential account may be seen in Adorno's *Dialectic of Enlightenment* (1944), which he co-authored with Max Horkheimer. According to Rose's account of *Dialectic of Enlightenment*, alienation specific to capitalism is discarded in favour of a more trans-historical attention account undermining it as an account of capitalism (Rose 8).

Rose's argument that the first cultural theorists overlooked those aspects of Marx's writings which pertain to class struggle and the extraction of surplus value, in favour of an idealistic cultural criticism are germane in accounting for why it is that works such as Eagleton's *Exiles and Emigrés* (1970), Williams' 'When Was Modernism' (1987) and Jameson's *A Singular Modernity* (2002) regard the agency of the revolutionary subject as playing the same role Spirit did for Hegel (Callinicos 92; Eagleton, *The Ideology of the Aesthetic* 225). The aesthetic is furthermore foregrounded as though it were the key determinant of collective agency. Despite the greater degree of attention these critics aim to grant to historical causality in literary form, under this rubric modernism remains a paradigmatic instance due its correspondence in time with a heightened and protracted phase of class struggle on an international basis. We see this expressed most clearly in Perry Anderson's account of social revolution as one of the three most significant influences on modernist literature (Anderson, "Modernity and Revolution" 104). The other two include the abiding of the European *ancien régimes* within industrial modernity, as outlined by Arno Mayer in *The Persistence of the Old Regime* (1981), and the transformations wrought by rapidly developing communications technology. Anderson's contribution to this debate is prompted by the publication of Marshall Berman's study of cultural production within both modernity and post-modernity, *All That is Solid Melts into Air* (1982). In a review of the book in *New Left Review*, Anderson charges Berman's text with promoting the idea that modernist literature proceeds more or less in lock step with processes of economic and political modernisation; the development of productive forces and culture are identified as proceeding continuously in a linear direction as the twentieth century advances (Berman 90–97), more in line with a Weberian than an orthodox Marxist point of view. Williams, by contrast, emphasises changes in institutions of cultural production, especially technological change (Williams, "When Was Modernism?" 50). Jameson, meanwhile, extending the philosophical legacy of critics such as Jürgen Habermas and Adorno, emphasises an anti-systemic or negative perspective on Hegel's philosophy (Dews 125–26) introducing a greater degree of conceptual flexibility to the notion of the aesthetic transformation, arguing that a break may be prolonged, overlap with other periods or constitute a period in its own right (Frederic Jameson, *A Singular Modernity: Essay on the Ontology of the Present* 23).

On whatever terms one may seek to criticise any of these models briefly outlined above, their capacity to incorporate both past and future accounts should not be overlooked. Any of the thousands of studies written on the topic of modernism's genesis which seek to account for why there was a general growth in the sentiment that the then-established means of representing reality had become increasingly inadequate will take some facet of one or more of these three as the decisive contradiction, to which some additional points may be introduced by way of content. Some of these include a generalised societal desanctification which accompanied a decline in Christian belief, catalysed by World War I as well as the growing influence of ideas derived from the natural sciences and German higher criticism (Bell 10–11; Blair 164–65; Said xxix; Williams, "When Was Modernism?" 50). Other critics working in a more directly historical material, identify the effects of these revolutionising changes within the sphere of cultural production, with industrialised printing allowing for the production of large numbers of newspapers cheaply, and this, coupled with an increasingly literate public and the development of mass communications, provided the infrastructural supports through which serially published novels could begin to challenge the cultural hegemony of the multi-volume realist novel, especially once Britain's oligarchic subscription libraries had entered into a sustained financial decline (Caldwell 16–17; Childers 77–78; Cleary, *Outrageous Fortune: Capital and Culture in Modern Ireland* 71–72; David 5). Lawrence Rainey in particular, has drawn attention to the role luxury book speculation and networks of patronage played in providing many of the most prominent literary modernists with the means of sustaining themselves (Rainey).

Given the degree to which historical events and development have been incorporated into the history of literary criticism, it can be difficult to understand how it is that the break away from the Hegelian idealist dialectic has not been accomplished. The first and most significant reason could be associated with the Western Marxist tradition's turning away from questions of class struggle from the mid twentieth-century onwards and more towards issues associated with culture and artistic representation, a process which Anderson identifies (Anderson, *Considerations on Western Marxism* 75–76). Based on the history of international class struggle especially from the seventies onwards, the dialectic comes to be increasingly formulated as a tragic framework, more appropriate to the crises of representation and language of post-modernity than enlightenment teleology (Fredric Jameson, *Valences of the Dialectic* 3; Frederic Jameson, *The Hegel Variations:*

On the Phenomenology of Spirit). Considered as such, Hegel does not provide any means of unifying positive phenomena or understanding human behaviour within a totality, so much as he complicates our object of study (Fredric Jameson, *Valences of the Dialectic* 31). As we have already seen, Hegel likewise regards Spirit as developing through a constant return to itself and encompassing all of its moments sequentially; over time in a motion which may be described as circular. In this way, Spirit represents an organic development out of its own concept (Hegel 24). Hegel's philosophical system therefore insists upon a distinction between two distinct types of teleology; Hegel's circle returns to itself at progressively higher levels of sophistication and complexity, rejecting the 'bad infinity' of the straight line which ascends in a perpetual linearity (Knox), the 'bad infinite' which we might compare to Anderson's criticism of Berman's conception of cultural change within modernity. This could be part of the reason why the synonymy of political and aesthetic progress is assumed within Anglo-American Marxian analysis, a major weakness of the school which Sinéad Kennedy identifies (Kennedy 127). Eysteinsson locates the cause for this in the lack of an English-speaking *avant-garde* tradition. This has the consequence that the Anglosphere lacks a significant aesthetic movement, which in continental literary criticism, provides the negative riposte to institutional or classical forms of modernism (Eysteinsson 85).

It is necessary at this point to argue that this emphasis on complexity and further periodisation over the past half-century has in many ways been immensely productive. Some of its legacies include the more inclusive models of literary history produced by Sandra Gilbert and Susan Gubar or Nancy Fraser, Edward Said's studies of the literature of imperialism, as well as the general incorporation of developments within the visual arts, film, fashion, architecture, scholarship and dance. The more cultural studies inflected schools of literary criticism such as neo-victorianism or new modernist studies, which have all been decisive in challenging the value-laden assumptions regarding the distinctions between high and low art (Mao and Walkowitz, "Introduction"; North 91) represent its contemporary iteration. However, with the emphasis on the mediations between history and literary content, it becomes increasingly difficult to retain our focus on any literary genre in and of themselves as they are made to mediate an increasing and overdetermined sequence of agents of revolution and social change. We see this in the proliferation of categories associated with modernism over the past few decades as in 'late modernism' (N. Allen; MacKay 15–16), 'post-war' modernism (Mengham 71) or 'cold modernism' (Burstein). Experimental literature of the early twentieth century therefore remains for all intents and purposes the horizon of literary and aesthetic achievement, to the extent that realism, the gothic and sensational literature are all subject to potential re-readings as modernist modes (Flint; David) and 'modernists' such as Joseph Conrad or Kate Chopin are rendered as far more proximate to traditions within which such predecessors as Harriet Beecher Stowe or Elizabeth Gaskell might be placed (Frederic Jameson, *The Political Unconscious: Narrative as a Socially Symbolic Act* 192–212; Kalaidjian 3–4). This culminates in a situation wherein the co-ordinates of literary history are consistently up for debate, as the old is taken to bear the impress of the new and what was formerly regarded as the new is increasingly sublimated to an abiding traditionalism. Emily Apter's presentation of provisional, asynchronous, de-sequenced and localised models of literary history oriented in activist directions as the superior alternative to totalising approaches contaminated by their proximity to capitalist and imperialist logics of domination, may be read as symptomatic of this tendency (Apter 3, 65). In this sense, literary criticism becomes itself a modernist project, steered primarily by the impulse to 'make it new' or overcome the conceptual reification of categories, rather than to solve or contextualise in a way which would clarify rather than facilitate further argument (Fredric Jameson, *Valences of the Dialectic* 51). One can see how poorly an authentically Marxian theory of literary history, as well as one steered by empirical approaches, with its emphasis on historical determination would fare within a literary-critical milieu within which contingency has become hegemonic. Though as Firdous Azim notes, literary criticism has been conceptually mobile throughout its history and adapting critical concerns to the imperatives of contemporary political commitments lends many works of criticism their particular novelty and impetus (Azim 243), the peculiarities of the predominance of theory and cultural studies in the humanities, as traced by Joseph North, now takes place at a point in the history of the humanities which is qualitatively distinct (North 10, 57). This is not solely due to the ways in which these ideas themselves have acquired their own momentum. In much the same way as we have already attributed the professionalisation of literary criticism to the construction of the modern capitalist state, we can identify the greater amounts of interest associated with cultural as opposed to historical concerns within literary criticism as being consonant with the broader logic of third-level education in Ireland as well as other jurisdictions at the present time. This can be illustrated by referring to indicative literature produced both by the Irish government and research-oriented consortia located both within the Irish state and the European union.

Within this literature, we see the foregrounding of the imperatives of the post-industrial 'knowledge economy,' wherein western states intend to remain competitive internationally by incentivising universities to prioritise research which aligns with the interests of private enterprise, such as biotechnology, information technology and financial services, according to a tendency which Kieran Allen has outlined, whereby in an age of globalised capital, the objectives of industry and universities are increasingly difficult to differentiate from one another (K. Allen 135–38). When we investigate the figure of the discipline-specific department in this literature, we see it consistently invoked as a fetter on the development of an authentically interdisciplinary research environment, due to its tendency to inhibit the movement of knowledge across disciplinary boundaries, rendering them inadequate to tackling problems held to be of global significance, such as climate change or 'migrant crises' (Wernli and Darbellay 5–7). This is due to their 'inwardly directed social dynamics' and the fact that the frameworks through which departments assess validity originate from within their own disciplines rather than from without (Mazzucato 12–15; Wernli and Darbellay 17). This makes the task of any supra-departmental management body seeking to implement quality assurance standards from above more difficult, a mechanism which is increasingly to the fore within university administration, as documented by Wendy Brown and Stefan Collini (Brown 23; Collini 1–2). One can see how interdisciplinary research potentially offering the opportunity to outflank the comparatively rigid autonomous department would represent an attractive prospect to university administrators (Readings 39) and it is in this context that the EU's primary funding instrument placing renewed emphasis on interdisciplinary research should be comprehended (Department of Jobs, Enterprise and Innovation; Wernli and Darbellay 9). In accounting for these changes we would not wish to merely echo North's argument that by virtue of their coinciding at a point in time when the university has been transformed into an institution run in the interests of profit that theory-driven or post-Marxist histories of literature are inveterately neoliberal; it is important to understand instead that they merely manifest the economic and financial incentives which are a function of the contemporary university's primary stakeholders, especially when this material may maximise a particular research project's public impact. Though such policy literature is overwhelmingly directed towards universities' STEM outputs — it being where the more straightforwardly commercialisable research sectors are located — it is by no means clear that literary studies can be viewed apart from these broader logics. Just as Fordism was influential in shaping the institutionalisation of literary criticism as we saw above, the current precarious funding environment inevitably influences the type of research being produced. Both Max Brzezinski and Charles Altieri have for example, undertaken close readings of contemporary culturalist readings of literary genre as terminating in aestheticism, effacing differences between actually-existing schools of political thought, simplifying context and subsuming reactionary, liberal and left-wing traditions of literature and literary criticism within single categories, terminating in anachronism, the over-writing of political differences, or both (Altieri; Brzezinski 113–14). In this sense a predominantly culturalist approach to literary studies ironically comes to reproduce once again the logic which it ostensibly begins by attempting to reject, reproducing the value judgements of modernist form, the impasse which led to the attempted Marxian extrusion from new criticism in the first place.

4.4 Finding words

Before beginning this account it will first be necessary to outline how we came to the specific literary instances which are dealt with below. As we saw in *Figs.* 3.3, 3.4 and 3.5, there are no data points which can be identified as being especially high for positively correlating words, the pattern is too incremental to identify a single breakthrough year, either through heuristic or mathematical means. This problem is compounded with regard to the negatively correlating words based on how much more restrictive these trend lines are. At first we attempted to identify texts from the beginning and end of the trend line, finding, for example, novels published in the 1750s which scored particularly highly for negatively correlating words and texts published in the twentieth century for high amounts of positively correlating words. This approach was unsatisfactory for two reasons. Firstly because it turned out that reliable digital transcriptions of most texts contained within the HathiTrust word frequencies collection were exceptionally difficult to obtain. Transcriptions that were available on The Internet Archive were of such a low quality that they were effectively unusable for our purposes, even with Underwood's contextual corrections provided on his GitHub page (Underwood, *DataMunging:rulesets at master · tedunderwood:DataMunging · GitHub*). This problem is partly associated with the general unavailability of digital versions of literary texts published or written before the nineteenth century detailed in the previous chapter. The second reason this solution was unsatisfactory is that there is a risk that it would leave us with no points of comparison. If we were to take negative correlations from early in the dataset's history and positively correlating from late it gives little sense of a transition from one state to another, but rather remains somewhat static. On the basis of the fact that incremental rather than sudden change is what we confront in our findings, we decided to take the positively and negatively correlating words which we have extracted and to identify particular sections of texts which have been afforded some degree of significance within the literary historiographies proposed by some of the critics we have already considered. We identified specific sections of these texts which are particularly dense with either positively or negatively correlating words by loading these texts into the R workspace and dividing them into 150 word chunks. The degree to which each of these chunks were composed of either positively or negatively correlating words was calculated and those which scored particularly highly were extracted, under the assumption that these passages would be adequate to the task of providing an insight into changes taking place in literary history over time. The words which correlate either postively or negatively to literary change in each instance are highlighted in bold, allowing the reader to identify tangible instances of these changes at their most pronounced.

4.5 Realism in literature

We begin our consideration of the literature itself with an extract from Charles Baudelaire's poem 'The Voyage,' translated from French by F.P. Sturm and published in 1919.

But when at length the Slayer treads **us** low,
We will have hope **and** cry, "'Tis **time** to go!"
As when of **old** we parted **for** Cathay
With wind-blown **hair and eyes upon** the bay.

We will embark **upon** the Shadowy **Sea**,
Like youthful wanderers **for** the first **time free**
Hear you the lovely **and** funereal voice
That sings: O **come all** ye whose wandering joys
Are set upon the scented Lotus flower,
For here we sell the fruit's miraculous boon;
Come ye **and** drink the **sweet and** sleepy power
Of the enchanted, endless afternoon.

VIII

O Death, **old** Captain, **it is time**, put forth!
We have grown weary of the gloomy north;
Though sea **and sky** are black **as** ink, lift sail!
Our hearts are full of **light and will not** fail.

O pour thy sleepy poison in the cup!
The fire **within** the **heart so** burns **us up**
That we **would** wander Hell **and** Heaven **through**,
Deep in the Unknown seeking **something** new! (Baudelaire, *The Poems and Prose Poems of Charles Baudelaire*)

‘The Voyage’ is a poem comprised of thirty-seven stanzas divided into eight sections. Our method focuses our attention on the final four stanzas. The poem is an ironic re-writing of the Homeric epic and could be regarded as a satirical treatment of epic poetry in general. Throughout the poem, we see consistent parallels being drawn between the ordinariness of modern life, which is rendered primarily by reference to the boredom of domesticity and one’s home country in general. ‘The Voyage’ is therefore concerned with rendering the ambivalence between these two distinct modes of existence, with particular emphasis laid upon the way in which the epic journey abroad represents the opposite of normal daily life, especially it seems, insofar as its promising to offer something like the experience of the exotic or the Orient. There are suggestions earlier in the text of the madness or death which may await those who seek such adventures out and it is in the four stanzas above that we begin to see these two modes of existence initially presented as opposites begin to be threaded together. This is accomplished primarily by the figure of the Slayer, who we understand as being symbolic of death, which is itself formulated as a realm much like the far East, in its capacity to offer both material and spiritual satisfaction or indulgence, to the extent that they become more or less undifferentiated. This coming together of the two opposites, based on the fact that both end in death, can be seen as being reflected on the typographical level. While earlier in the poem there are quotation marks which signify the points at which the speaker addresses an auditor, by this stage in the text these quotation marks have fallen away, indicative of the subsumption of one into the other.

Before locating this poem in its historical context, it is necessary to note that this text is an excerpt from a poem and while the computational literary-critical discourse is comparatively rich with quantitative accounts of novelistic history, there has been almost nothing written in the way of a quantitative history of poetry. By itself, this thesis will not bring about any significant change in this state of affairs, but it is worth highlighting that the results both arrived at in the previous chapter and detailed in this chapter, attest to the fact that the changes which occur in both poetry and prose over the roughly two-century time period this study considers are reasonably analogous up to a point. We see 121 words positively correlating words shared between the two genres and these are words which are associated with physical appearances (‘hair,’ ‘eyes,’ ‘face’) as well as concrete nouns associated with descriptions of the physical environment (‘door,’ ‘grass,’ ‘sky’). The remaining 58% of words which appear in the poetry dataset are words which are effusive, emotive (‘dear,’ ‘love,’ ‘longing’) religious (‘angel,’ ‘god’) or refer to the natural world (‘birds,’ ‘lilies,’ ‘sea’). Poetic language therefore maintains a focus on social relationships, whether friendly, familial, romantic or devotional in a manner that fiction does not. There is also a significantly greater amount of words associated with emotions and sensibility; we see a corresponding decrease of these in fiction over time, which we see in Appendix E and in Underwood’s findings (Underwood, *Distant Horizons: Digital Evidence and Literary Change* 25). In this sense, poetry and fiction seem to accord to slightly different trajectories, with both becoming more focused on outward appearances, with one retaining an interest in introverted subjective feeling simultaneously. The most relevant words here, insofar as our quantitative procedure is concerned, seem to do with the short, clear and specific descriptive writing associated with parts of the body and positive adjectives as well as concrete description. For the moment though, we will concentrate on the action which takes place between two figures described in 'The Voyage'; an unnamed speaker rendering an impressionistic history of voyages they have undertaken in the past to an unnamed auditor. In situating this in its context, we will now consider Baudelaire's presence at the outset of many histories of modern literature, due to the emphasis his works place on the modes of cognitive mapping particular to the individual and artist within the modern environment.

Fundamental to Baudelaire's conception of the artist within modernity, as outlined in his essay 'The Painter of Modern Life' (1863), was precisely this attunement to extreme emotional states, especially when they are ephemeral in nature (Baudelaire, *The Painter of Modern Life* 2). This is very much to the fore in the closing stanzas of 'The Voyage' with comparisons made between the death of the Greek philosopher Socrates and the Christian afterlife lending a dream-like sense of fluidity and spontaneity to affairs as they occur. This bringing together of the Christian afterlife and Ancient Greek history stages a synecdoche of the poem in general, a repudiation of real life, in favour of the cognitive, sensory or emotional sensibilities which respond to, mediate and perhaps even supersede this 'real.' This holding up of the transcendent or subjective over and above the mundane represents a trope which has been identified within some accounts of modern literary art (Habib 562). As Berman notes 'Modernism, then, was the quest for the pure, self-referential art object…the proper relationship of modern art to social life was no relationship at all' (Berman 30). Eric Hobsbawm's description of aestheticism, the literary movement steered via the maxim of 'art for art's sake' with which Baudelaire is often identified, accords with Hegel's account of romanticism, regarding it as being symptomatic of an inward turn in European intellectual and cultural production, a retreat into formalism and self-consciousness after the failed European revolutions of 1848. Implicit within this account of aestheticism's origin is the suggestion that aestheticism represents the rejection of any aspiration to societal critique (Hobsbawm 325). From what we have seen so far, Baudelaire's text focuses to a greater extent on the local, individual or imaginative experience at the expense of a particular context, which is supplanted by a far broader frame of imaginative and geographical reference. There is also perhaps more than a little nostalgia for the pre-urban or pre-Christian sensibility in the adoption of the epic as a governing symbol, a nostalgia which Eagleton identifies as crucial to modernism's dependence on myth (Eagleton, *The Ideology of the Aesthetic* 60). It is not until the final verse of the poem that what we might consider the poem's sublimated conventionality is revealed. The speaker, representing the coming together of the voyager and the auditor, addresses death and announces their determination to sail from the gloomy north, which may represent the 'northern continent' of Europe, though of course equally, we might read these final lines as the introduction of yet another layer of subjectivity or imaginative fancy instead of a more precise localisation.

Having been examined both from the perspective of its form as well as its content, Baudelaire's text seems to convey the non-specificity or estrangement of the milieu being contained within a language which we can see is increasingly characterised by descriptive specificity. Instances of emotion which appear in 'The Voyage,' oriented around resolve, discovery, pursuit of the new, function autonomously or in isolation from a precise referent, it seems, precisely for this reason. It may be within the highlighted past and present participles that we see this contradiction reach its apex, wherein a narrative structure, the differentiation between a past and present is beginning to enter into poetry at the same time that a clear or unambiguous account of its precise situation is beginning to reach its vanishing point (Butler 81; McFarlane 82). The question we must now consider is whether or not the innovations we have attributed to Baudelaire here are entirely without precedent. As we have already suggested, the affirmation of individual sensibility and vision is hardly a novel invention given the legacy of the romantic artist within nineteenth century literary culture (Williams, *Culture and Society 1780 - 1950* 56; Wilson 4). It is for this reason that we now consider the last two verses of William Wordsworth's poem 'The Mad Mother,' from the poetry collection Wordsworth co-authored with Samuel Taylor

Coleridge, *Lyrical Ballads* (1798):

Oh! love me, love me, little boy!
Thou art thy mother's only joy;
And do not **dread** the waves below,
When o'er the sea-rock's edge we go;
The high crag cannot work me harm,
Nor leaping torrents when they howl;
The babe I carry on my **arm**,
He saves for me my precious soul;
Then happy lie, for blest am I;
Without me my sweet babe would die.

Then do not fear, my boy! for thee
Bold as a lion I will be;
And I will always be thy guide,
Through hollow snows and rivers wide.
I'll build an Indian bower; I know
The leaves that make the softest bed:
And if from me thou wilt not go,
But still be true 'till I am dead,
My pretty thing! then thou shalt sing,
As merry as the birds in spring (Wordsworth and Coleridge).

It is important that in outlining the crucial and quite distinct differences between Wordsworth and Baudelaire's text that we do not implicitly argue that Wordsworth's poem is less competently executed than Baudelaire's or that it is without ambivalence or ambiguity. Having said this, the impression of the speaker which the reader is invited to take away from 'The Mad Mother,' is emphasised to a far greater extent than in Baudelaire. Though there are far fewer negatively correlating words in our poetry database and therefore much less to work from in terms of secure differences, it can be said that there is a certain degree of simplicity in the three words we do have ('dread,' 'nor' and 'arm') in comparison to some of the word-types we considered in Baudelaire. The differences begin as early as each works' respective titles, Wordsworth offering far more in the way of an objective signposting than Baudelaire. In addition to this, the symptoms of modernity we see in Baudelaire, conquest, exploration, myth, a spurned domesticity are nowhere to be seen in Wordsworth, having been replaced by imagery relating to the natural world, rendering them far more familiar or oriented in and of themselves. Within Baudelaire, there is, as we would expect, a certain degree of ambivalence regarding the substantiality of modernity, while in Wordsworth, the crag, the river, the landscape all locate us within a natural milieu which is far more clearly representative of avenues into freedom, individuality and autonomy from society, all libertarian themes which would be familiar within Wordsworth's *oeuvre* in general (Connell). Both texts involve or require the subsumption within the speaker's perspective, an insistence on the individual point of view, although as signalled earlier, Baudelaire's speaker seems to be composed of more than one individual brought together at the poem's conclusion, while in Wordsworth, the speaker is in view from the poem's outset. Baudelaire's text abdicates to a greater or lesser extent many of the animating principles which lie behind poetic form, whereas within Wordsworth's text, the poem's objectives are accomplished in part by having the form geared towards this objective; the placement of the punctuation encourages a reading which expresses the instability or distracted temper of the speaker, the inconsistent arrangement of the half or rudimentary full-rhymes, gesture towards a central unresolvedness. This effect is offset somewhat by the use of the language, which seems far more gentile than one might expect within a poem voiced by someone who is mentally ill; not even the regional accent required to bring some rhymes to their conclusion can quite accommodate this dissonance. It would be a relatively straightforward matter to trace this growth of expressive form in Wordsworth being brought to its point of collapse in Baudelaire; Malcolm Bradbury and James McFarlane are among the many critics who have read the deterioration of the myth of the individual as significant to any consideration of modernist literature (Bradbury and McFarlane 26).

Given the continuities we have outlined between romanticism and modernism, it may be worthwhile to grant more close attention to its definition. Romanticism is described by Terry Pinkard as one of the imaginative and cultural responses to the French Revolution, a means of locating the subjective experience of the individual relative to broader discourses associated with political liberty and emancipation from monarchical authority. This manifests itself most clearly in the breaks from classical form, which Abrams has identified, but also in a manner which is less overtly obvious. Pinkard for instance, identifies the historical unprecedentedness of the centrality afforded to the individual imagination within European modernity as a distinctive aspect of cultural expression which emerges in the nineteenth century (Pinkard 96–97). Hegel, writing on this key juncture within the history of Spirit, identifies its arrival as indicative of the movement of Spirit more in the direction of a self-undermining scepticism, a period of time in which previously regnant ideals are dissolved (Pinkard 208–10). Art produced within the context of European modernity, will therefore inevitably exhibit a greater tendency towards inwardness and individuality, with the result that the critical reception of works within which the individual's concerns are emphasised; Pinkard identifies Shakespeare's greatness as a dramatist being secured in the nineteenth century precisely due to Shakespeare's characters individually focalised objectives as a suggestive development in this context. This has the consequence, however, that art no longer manifests a devotion to any broader Idea, such as God, society or the state, except as a partially ironised exercise. Modern art then becomes an exercise within which the subject exhibits themselves to the viewer, in a recognition of a lack of unmediated access to these 'givens' (Pinkard 600–03). Eagleton has made these trends, described above, legible within the context of literary modernism, identifying Hegel's conception of Spirit with the alienated subjectivity of the bourgeois subject (Eagleton, *The Ideology of the Aesthetic* 123). Habermas does so specifically with regard to the works of Baudelaire, describing how, as the process of modernisation begins to accelerate in the nineteenth century, established means of locating oneself relative to one's environment becomes increasingly untenable (Habermas 8–9). It is here that we may once more consider the potency of Wordsworth's conception of nature, in contrast to Baudelaire's admixture of Pagan and Christian mythology as if in pursuit of some kind of symbolic framework that can make for up for the absent Idea.

So far we have identified the continuities between the romantic sensibility at the outset of the nineteenth century and the putatively more modern *geist* in and around its halfway point. We have pinpointed the through line common to these two moments in the emphasis which is placed on the figure of the individual and their subjectivity in order to demonstrate that any previously existing *terra firma* within which they might have been described or located in the past no longer exists. This excision is concurrent with an increasing focus on the concretely described object, developments which would be difficult to consider apart from the processes indicative of capitalism and modernisation, such as the growth of the industrial economy and the secularisation of moral values and beliefs. We will now investigate this phenomena in greater amounts of depth and begin to introduce some of the more securely Marxian aspects into our interpretative framework.

4.6 Naturalism in literature

It is in the reading of literary genre known as naturalism that the robustness of a Marxian literary critical approach as it has been conducted within the context of the New Left is most evident. This is due to the degree to which it can be associated with the rise and development of modern state governance and the actuality of governing a population, stratified along class lines and often concentrated within the urban environment (Frederic Jameson, *The Political Unconscious: Narrative as a Socially Symbolic Act* 120). Williams identifies one of naturalism's key tropes, namely, the registering of the existence of the various disciplinary regimes of knowledge production which have arisen in order to accomplish this end, such as criminology, neurology and psychiatry in the sphere of literature Williams (*Modern Tragedy*). Joe Cleary, contextualising the milieu from which naturalism emerges in more specific terms, writes the following:

> This intellectual climate helped to mould the naturalist conception of the writer, articulated most famously by Zola in his prefaces and manifestos, as a detached, clinically objective 'scientist' of human nature or society, with a duty, like that of the scientist or doctor, to vivisect the tissue of conventional moral niceties in pursuit of the deeper 'laws' that governed human behaviour. This emphasis on scientific objectivity, and the conception of the novel as a laboratory where experiments concerning individual and social behaviour could be conducted, contributed to the much commented upon determinist sensibility that supposedly characterises naturalist fiction: its assumption that the laws of heredity and social environment, abetted by the undersell of an ungovernable sexual instinct, allowed for only a very constricted form of human agency (Cleary, *Outrageous Fortune: Capital and Culture in Modern Ireland* 113–14).

Naturalism's 'view from nowhere,' focalised via the viewpoint of a particularly forensic narrator, is therefore aligned with the interests of state power, which accounts for naturalism's 'scientific' emphasis on societal dysfunction, legible from a contemporary perspective as social crises denotative of modernity, such as alcoholism, adultery or violence, themes which are present to a significant extent in Crane's fiction, which we consider further below.

The literary critic Toril Moi provides what is possibly the most coherent means of locating realism and its derivates on a trajectory within the development of modernism and on this basis more worthy of considered critical assessment. Moi hypothesises that realism, naturalism and modernism all form part of what is essentially the same trajectory in modern literary production; a shared commitment to the rejection of idealism in artistic representation. Idealism is here defined as a didactic impulse within literature, an effort on the part of the author to morally improve or to cultivate the reader, an ethic which Moi regards as increasingly marginal to literary production within modernity (Moi 67). This forms a contrast with previously regnant new critical stereotypes of naturalism as a formally inept literary genre, culminating in the current situation, where much of the critical discourse surrounding naturalism is one characterised by reclamation and recovery, re-contextualising authors such as George Gissing, Thomas Hardy, Jack London, W. Somerset Maugham and Arnold Bennett as significant within the history of literature (Joyce, *Modernism and Naturalism in British and Irish Fiction, 1880–1930* 8–13). In an Irish context too, we see an increasing degree of attention afforded to the works of George Moore, as a means of steering between the stereotyped antinomies of Irish revivalism or modernism (Nolan 172). It is Simon Joyce who takes Moi's greater sense of a continuous historical perspective still further, locating the origin of high modernist narrative within the writing of naturalist authors such as Moore, Stephen Crane, Honoré de Balzac or Émile Zola, reading the genre as an internationalist proto-modernism and far more *engagé* than high modernist impressionism against which it is typically compared in unfavourable terms (Joyce, *Modernism and Naturalism in British and Irish Fiction, 1880–1930* 9; Pizer 12–14). Moi's hypothesis offers the advantage of allowing us to regard these genres as forming an organic unity or, to paraphrase Michael Bell (Bell 9), different stages in the digestion of what is effectively the same formal impulse (Moi 9). It is then, in part potentially due to naturalism's influence, that we see the decline in words associated with religious terminology, in favour of a more mechanistic perspective and this indeed accounts for why it is that our results show is a surprisingly unified realism. Other genres held to have arisen in the period of time we are considering, whether realist, modernist or intermediary modes such as late realism, naturalism, impressionism or imagism, do not seem to take place at a scale which is detectable at the level at which we are studying. These categories are without doubt productive. An acknowledgement of realism's predominance does not render these stylistic or periodising terms irrelevant, but it may point to the difficulty that critics who relate literary discourse to broader historical and political developments have in recognising the generic nature of realism in its concrete manifestations. The fact that literature only seems to undergo one significant transformation from 1750 — 1922, and that this change is concurrent with the arrival and consolidation of realism therefore requires us to re-locate realism within its historical context. In developing these points further we might consider the work of Stephen Crane, an American writer who has been read both as a naturalist and an impressionist. Crane makes for a productive case study within this trajectory due to his being situated in two different camps at once. Martin Scofield describes how the occasion of the American civil war prompted Crane to move beyond the objectivity which characterises works such as *Maggie: A Girl of the Streets* (1893) towards *The Red Badge of Courage* (1895), which, in its representation of reality as existing in an intersubjective state of flux seems to anticipate the vacillating modalities of impressionist narrative discourse (Scofield 72). We might further consider Crane's status as an intermediary figure within this chronology by reading extracts from these two novels closely. First, we consider an extract from *Maggie: A Girl of the Streets*:

> The girl thought the arrogance and granite-heartedness of the magnate of the play was very accurately drawn. She echoed the maledictions that the occupants of the gallery showered on **this** individual when his lines compelled him to **expose** his **extreme** selfishness.
>
> Shady persons in the audience revolted from the pictured villainy of the drama. With untiring **zeal** they hissed **vice** and applauded **virtue**. Unmistakably bad men evinced an apparently **sincere admiration** for virtue.
>
> The loud gallery was overwhelmingly with the **unfortunate** and the oppressed. They **encouraged** the struggling hero with cries, and jeered the

> villain, hooting and calling **attention** to his whiskers. When anybody died in the pale-green snow storms, the gallery mourned. They sought out the painted **misery** and hugged it as akin.
>
> In the hero's erratic march from poverty in the first act, to wealth and triumph in the final one, in which he forgives all the **enemies** that he has left, he was **assisted** by the gallery, which applauded his **generous** and noble **sentiments** and confounded the speeches of his opponents by making irrelevant but very sharp remarks. Those actors who were cursed with villainy parts were confronted at **every** turn by the gallery. If one of them **rendered** lines containing the **most** subtile distinctions between right and wrong, the gallery was **immediately** aware if the actor meant wickedness, and denounced him accordingly (Crane, *Maggie: A Girl of the Streets*).

Secondly, we consider an extract from *The Red Badge of Courage*:

> His **tall** figure stretched itself to its **full** height. **There** was a **slight** rending **sound**. **Then** it began to swing forward, **slow** and **straight**, in the manner of a falling tree. A swift muscular contortion made the left **shoulder** strike the ground **first**.
>
> The body **seemed** to bounce a **little way** from the earth. "God!" **said** the tattered soldier.
>
> The youth had **watched**, spellbound, this ceremony at the place of meeting. His **face** had been twisted into an **expression** of every agony he had imagined for his friend.
>
> He now **sprang** to his **feet** and, **going** closer, gazed upon the pastelike **face**. The mouth was **open** and the **teeth** showed in a **laugh**.
>
> As the flap of the **blue** jacket fell **away** from the body, he could **see** that the **side looked** as if it had been chewed by wolves.

> The youth **turned**, with sudden, livid rage, **toward** the battlefield. He **shook** his fist. He seemed **about** to deliver a philippic.
>
> “Hell–”
>
> The **red** sun was pasted in the **sky like** a wafer (Crane, *The Red Badge of Courage: An Episode of the American Civil War*).

The first paragraph describes the occasion of a play which Maggie attends. While this play is described in the context of a number of plays and other spectacles that Maggie is brought to by a man named Pete, this particular paragraph seems to describe a specific play or performance. This attempt to render subjective experience within a broader continuity of events seems to terminate in a certain degree of ambiguity, the villains of the piece and the reaction of the audience at once mark specific receptions or occasions and a broader continuity of reactions to other plays which seem more or less similar. The overwhelming majority of negatively correlating word types seem to relate to moral judgements or moralising language (‘affections,’ ‘barbarous,’ ‘benevolence’), particularly regarding the nefarious behaviour of the villain, the positive qualities of the good characters and the sincerity with which the audience’s reactions are invested. The attempt this passage makes to summon up a more generalised portrait of previous productions means that the exact nature of the evil or goodness onstage eludes precise narrative description. It is furthermore important to note that this passage identified by our automated method describes a slightly retrograde form of populist art in which good and bad characters are easily identifiable There is a significant amount of irony in Crane’s representing this within a novel attempting to convey a more authentic portrait of modern urban life. The awkwardness of some of the syntax perhaps points towards the inappropriateness of these concepts in a modern context.

This all stands in quite stark contrast to the second paragraph from *The Red Badge of Courage* which describes an encounter between two soldiers, the novel’s protagonist, Henry Fleming, and Jim Conklin, who has been mortally wounded in battle. The emphasis here is more on two individuals than the broader collective we see described in our excerpt from *Maggie: A Girl of the Streets.* As a result, rather than a broader account of the milieu, within which no single individual is afforded any more space or emphasis than any other, we see extensive amounts of detail being expounded on the physical appearance of these two men more or less in isolation. One of the most pronounced points of comparison can be identified by comparing each respective paragraph's treatment of physical appearance; while in the first we see a focus on abstract values, morality and emotion, in the second we see far more concrete physical detail. The body parts which are mentioned ('shoulder,' 'feet' and 'teeth') seem to have no broader significance beyond a commitment to representational felicity as such. The words around them, furthermore, are simple and do not introduce significant amounts of additional information or detail ('left,' 'tall'). It is also significant that this descriptive passage takes place at a remove from the battle itself, almost as though there is an effort being made to isolate these soldiers from a broader generality of figures and that this occurrence is set apart from the more extensive happening of the battle in and of itself. The view from nowhere we see in naturalism has almost vanished completely, but we might say that Crane's movement from objectivity to more generalised description, and a tendency towards subjectivity, the wound looking as though it had been inflicted by dogs for example, can be attributed to Crane's experiences of the American civil war, as if he found naturalism thereafter to give insufficient weight to the *qualia* of experience (Levenson 157–58).

So far this chapter has set out some of the problems involved in accommodating our empirical findings within a broader history of literary criticism which has by its own account, adopted a more avowedly historical orientation to its approach based on the degree to which it emphasises the radical separation between nineteenth and twentieth century literary art. In a bid to introduce a significantly greater amount of the Marxian dialectic into our approach, we embarked upon a series of close readings, beginning with particular stanzas of Baudelaire and Wordsworth; these served to emphasize how sensibility and subjectivity become uncoupled from a surrounding context and begin to be explored or expressed for their own sake. This process takes place even as a greater degree or proportion of the text is simultaneously taken up by nouns and concrete objects. We then saw, through a reading of two quite distinct paragraphs written by Crane only two years apart, how impressionist prose, in its ambition to project a social totality with the shortcomings of naturalism taken into account, paradoxically terminates in its reconstruction in vaguer or more ironised terms.

We might consider this a repudiation of naturalistic detail in favour of the individual perspective, although this may be at odds with the closing lines of the paragraph, which threaten to introduce an almost cosmic, or avowedly Christian setting for the description. We could see in the sun being cast as a wafer either as a successful re-sacralisation of subjective experience, a Jamesian symbol of a movement from the visible to the valuable, or equally we might regard it as parodic; the promised phillipic is not delivered and perhaps points more towards the breaking down of form in the process of its being constructed beyond the realm of the objective even, or especially, as the attempt is made to map it from a subjective point of view. Having viewed this process in motion in the work of Crane, this chapter will now consider some of the infrastructural factors which might allow us to ground some of the epiphenomena we have documented so far within a more long-term process.

4.7 Marx and literary criticism

Nevertheless, we have yet to offer a secure means of identifying these changes with the more macro scale. We will now accomplish this via materials offered by Marx in his key work of political economy, *Capital.* It is in the second and third volumes of *Capital* that Marx provides an account of the various concurrent cycles of capitalist production, beyond the familiar account of the productive process whereby a given worker is alienated from the value which they produce. This involves a further consideration of the larger-scale movements and processes that are brought into existence by capitalist competition in an expanding world market and a distinction between two different forms of capital, both fixed and circulating. The interrelation between the expansion of capitalist markets across the earth and a broader publication infrastructure, along with the uniformity of language and cultural expression has been well-documented, not only in the polemical writing of *The Communist Manifesto* (1848) but also by historians of print literature such as Lucien Febvre and Henri-Jean Martin (Febvre and Martin 319). In the most straightforward terms possible, the term circulating capital refers to all capital which is used up within the period of a single turnover, whereas fixed capital carries over from one turnover period into another. A good example of fixed capital might be a machine or the factory premises on which the machine is located, while the circulating capital would represent the machine's inputs, the component parts of the commodity which form the direct components of the productive process' final output. As Marx emphasises, this process of circulation is put under acute pressure at all times by the requirement that a profit be realised within the capitalist mode of production. The capitalist needs the commodity imbued with surplus value to be sold and to return a profit in a timely manner so that the productive process may be initiated again, albeit at a broader scale. If the commodity circulates unsold in the market, time is wasted during which the capitalist is not recouping any profits. This places a renewed emphasis on the spatial and temporal aspects of production, as it is now in the capitalist's interest to abbreviate the period of time between the end of the actual production process and the consumption of the product and introduces factors such as transport, retail and credit infrastructure to production as these can all play crucial roles in allowing for the realisation of profit on a more timely basis. This objective tendency towards the increasing velocity of productive and circulation processes is a significant part of the reason why capitalism is such a revolutionising mode of production as the capitalist is forced to invest consistently in new technologies to monitor and to discipline wage labourers to the greatest extent possible in order to remain competitive, resulting in an attenuation of the period of time between production and the realisation of profit, forcing the circulation of capital to as close to zero as possible and overcoming all temporal and spatial barriers insofar as they interfere (Marx 648–49).

This can be illuminated by examining two final paragraphs. The first is from John Galsworthy's *The Burning Spear* (1919):

> 'Stay, **my** friends!' he said; 'here in darkness we can see better the true proportions of **this** great question of free speech. There are some who contend that in a democracy **every opinion** should be heard; that, just because the good sense of the majority will ever lead the **country** into the right paths, the minority should be accorded full and fair expression, for they cannot deflect the **country**'s course, and because such expression acts as a healthful safety-valve. Moreover, they say there is no way of preventing the minority from speaking save that of force, which is **unworthy** of a majority, and the negation of what we are fighting for in **this** war. But I say, following the great leader-writers, that in a time of national **danger** nobody ought to say anything except what is in accord with the opinions of the majority; for only in **this** way can we **present** a front which will seem to be **united** to our common **enemies**. I say, and since I am the majority I must be in the right, that no one who disagrees with me must say anything if we are to save the **cause** of freedom and **humanity**. I deprecate **violence**, but I am thoroughly **determined** to stand no nonsense, and shall not hesitate to **suppress** by **every means** in the **power** of the majority — including, if need be, Prussian **measures** — **any** whisper from those misguided and unpatriotic persons whose so-called **principles induce** them to assert their right to have opinions of their own. This has ever been a free **country**, and they shall not imperil its freedom by their volubility and self-conceit.' Here Mr. Lavender paused for breath, and in the darkness a faint noise, as of a mouse scrattling at a wainscot, attracted his **attention**. 'Wonderful,' he thought, elated by the silence, 'that I should so have succeeded in riveting their **attention** as to be able to hear a mouse gnawing. I must have made a considerable impression.' And, fearing to spoil

> it by further speech, he set to work to grope his way round the chapel wall in the hope of coming to the door. He had gone but a little way when his outstretched hand came into contact with something warm, which shrank away with a squeal (Galsworthy).

And D.H. Lawrence's novella *The Captain's Doll* (1923):

> But she threw the pencil down, having no more interest in her writing. She wandered to where the large telescope **stood** near a farther **window**, and **stood** for some **minutes** with her **fingers** on the barrel, where it was a **little** brighter from his touching it. **Then** she **drifted restlessly back** to her **chair**. She had picked **up** her puppet when she heard him on the stairs. She **lifted** her **face** and **watched** as he entered.
>
> 'Hello, you **there**!' he **said quietly**, as he **closed** the **door behind** him. She **glanced** at him swiftly, but **did** not **move** or answer.
>
> He took **off** his **overcoat** with **quick**, **quiet** movements, and **went** to hang it **up** on the pegs. She heard his **step**, and **looked again**. He was **like** the doll, a **tall**, slender, **well**-bred man in uniform. When he **turned**, his **dark eyes seemed** very **wide open**. His **black hair** was **growing** grey at the temples — the **first touch**.
>
> She was **sewing** her doll. Without saying **anything**, he wheeled **round** the **chair** from the writing-**table**, so that he **sat** with his knees **almost** touching her. **Then** he **crossed** one leg **over** the other. He **wore** fine tartan socks. His ankles **seemed** slender and elegant, his brown shoes fitted as if they were part of him. For some moments he **watched** her as she sat **sewing**. The **light** fell on her **soft**, delicate **hair**, that was **full** of strands of gold and of tarnished gold and **shadow**. She **did** not look **up** (Lawrence).

These two paragraphs were arrived at via the same means as all previous paragraphs we have already considered and discussed, but the particular authors in this instance were chosen on the basis of their both appearing in Virginia Woolf's essay, 'Mr. Bennett and Mrs. Brown,' (1924) an account of two opposing tendencies in English fiction, as Woolf understood them, at the beginning of the twentieth century. It is in this essay that Woolf draws a distinction between authors such as Galsworthy and Lawrence. The first author, Gaslworthy is a 'materialist'; the second, Lawrence represents an alternative direction for English literature (H. Lee 405). Woolf argues that the failures of materialist authors are rooted in their inability to move beyond the documentation of external phenomena within a debased public sphere and towards the formulation of a purer form of literary expression oriented along psychologistic lines and here we may recall Baudelaire's own emphasis on the role individual psychology should play within modern literary production (Joyce, *The Victorians in the Rearview Mirror* 3 19–22; Oser 91–92). While it is important to allow for the fact that Woolf was writing polemically and the trends which we have the capacity to view in retrospect were still in their early stages at the time in which Woolf was writing, it would seem as though her basis for separating the two camps are slightly wide of the mark. For instance, it is in Galsworthy's paragraph that we see the refraction of modernity via individual psychology, not only because it is a paragraph taken from a chapter in which the novel's protagonist, John Lavender, addresses a political speech to a room which he does not realise has been empty for some time. It is therefore Galsworthy who troubles at the reality of Lavender's perspective, rendering it partially and allowing the degree to which he is ignored as where the parody of the passage thrives. In the partial terms in which Lavender's words are delivered we see the capacity for the individual viewpoint to supersede a violent and debased modern urban reality characterised by an anonymous rabble; there are some scuffles in the audience in the lead-up to Lavender delivering his speech. Lawrence's paragraph is also that of a particular character narrating what they see descriptively, in this instance Countess Johanna zu Rassentlow narrating mundane aspects of their daily routine. No transcending gestures occurs here, even the final sentence, 'She did not look up' actively repudiates any kind of movement beyond the immediately perceptible.

Jameson's writings on the distinct ways in which the body is articulated and represented in nineteenth as compared to twentieth century literature pertains directly to this question. Jameson argues that before the onset of bourgeois modernity, which he dates to the mid-nineteenth century, the body is more or less absent from literary history and when it, or other nouns, are employed they more often serve an allegorical function which symbolises faded respectability, displaced psychological functions or states, whereas Flaubert and Baudelaire introduce nouns and relationality between these nouns and especially in the context of sensory perception to literary production (Fredric Jameson, *The Ancients and the Postmoderns: On the Historicity of Forms* 31). Both Williams and Jameson provide accounts of modernist aestheticism which allow us to position Woolf's essay as the necessary obverse of this changing milieu; the insistence on a sector of society within which an autonomous role for art is being upheld or advocated for necessarily entails its commodification elsewhere (Frederic Jameson, *The Modernist Papers* 260). In accounting for why it might be the case that Woolf might still attribute a superiority to Lawrence's approach, while bearing in mind that it was not necessarily these specific works or paragraphs which Woolf would have had in her mind while writing her essay, we might consider the differences existing in each author's respective fields of vision. Galsworthy's attention, however disclaimed, is significantly rooted in the urban and public space, whereas the domestic sphere commands most of Lawrence's description.

We conclude this chapter in re-capitulating the material which this chapter engaged. We first laid out some of the difficulties involved in integrating an incrementalist account of literary history with the tradition of Marxian literary criticism, given its emphasis on the existence of clearly defined breaks in literary history, an inclination which it inherited from previously regnant schools of literary critique. These difficulties have, in recent years only become more pronounced as literary criticism becomes increasingly culturalist as opposed to historicist in orientation within the research environment of the neoliberal university system. We then proposed some points of departure for bringing together our empirical findings with a renewed Marxian emphasis by conducting a series of close readings of, in sequence, Baudelaire, Wordsworth, Crane, Galsworthy and Lawrence, relative to Marx's theories of spatio-temporal compression and crisis. In so doing we provided a secure means of periodising the literary genres of romanticism, naturalism and in so doing identify them as specific junctures within the incrementalist trajectory. A number of the analyses which we have outlined in doing so are by no means significantly removed from a classically Marxist critique. This demonstrates both the utility of dialectical materialism to literary study while emphasising the necessity of adequate calibration or adjustment of the arguments as appropriate. Our conclusions nevertheless remain somewhat provisional. This is due to the fact that the arguments we make here would be significantly enhanced by an expanded dataset or, in theory, a modified distance-based method which might better account for a more specific or concentrated cohort of words within which the lineaments of historical change might be discerned in a more focused manner. We therefore hope to have laid the foundation for a quantitative literary history which will look both to empirical methods as well as dialectical materialism in order to advance its critique, providing as it does the most secure means of locating objective entities within an historical process and the most coherent means of incorporating statistics within a historiography of literature. In concluding, we set this approach in opposition to one of Jameson's proposed models of literary change, wherein autonomous moments which takes precedence over any sense of historical development and the existence of a broader assemblage or totality is denied completely (Fredric Jameson, *The Ancients and the Postmoderns: On the Historicity of Forms*).

Conclusion

Structure of Thesis

In the first chapter of this thesis, a history was provided of the field of CLS from the mid sixties until the present day. This first chapter, which took the form of a literature review, was particularly focused on the methodological and theoretical developments within CLS and how the latter exerted a significant influence on the former. For instance, a certain number of early CLS scholars were particularly invested in CLS as a means of challenging the hegemony of social theory in the study of literature and regarded empirical approaches as a means of re-introducing the existence of the author as an unique entity with a singular and historically unique writing style, which locates its unity in the unconscious mind of the individual. This unconscious could be investigated by quantifying the use of particular formal features which may compose a stylistic fingerprint. Such methods were initially arrived at by CLS scholars such as Hoover and Burrows. As the increasing efficacy of CLS methods based on the quantification of high-frequency word-types became increasingly clear, the attempts to employ quantitative methods as a means of enlisting the unconscious mind against the death of the author became an increasingly fraught endeavour, especially once scholars such as Eder and Rybicki among others carried out a series of benchmark and optimisation analyses which took Burrows' Delta method as a starting point. Such studies demonstrated just how deep into word frequency strata the stylistic signature of the author could be identified, while significantly complicating the notion itself by introducing sociological categories such as language-use and time period into analyses, moving beyond the matter of individual authorship in isolation. This development began to cast a certain amount of doubt on the figure of the author as a conclusive and inviolable ground truth in accounting for clustering behaviour or differences between texts. Some attempt to account for the unfortunate fact that no significant amount of cross-pollination has ever taken place between CLS and social theory to an extent which might illuminate some of the tensions between these two distinct schools of thought was also accounted for in this chapter.

The second chapter was methdological and provided an account of the application of Jannidis et al.'s cosine distance-based improvement to Burrows' Delta method to each year in three literary corpora, each spanning a timeframe of approximately one hundred and seventy years. The cosine distance was calculated between the relative frequencies of each years' MFWs and a battery of *t*-tests were applied to the resultant distances in order to identify a specific period of time within which we would expect to see a qualitative transformation of the literary milieu. Once this figure had been identified, we calculated distances forward and backwards in order to identify years which might be said to constitute particularly influential agents of change within literary history.

The third chapter considered some of the results arising from a series of correlation tests applied to these distances as well as the application of a regularised logistic regression to years on either side of these breaks in order to identify whether or not a model could be trained in order to differentiate between a pre or a post break milieu. We then correlated relative word frequencies over time with the certainty of the model's judgement in an effort to identify which word-types were most indicative of these transformations. Once the relative incidence of these word-types had been plotted on graphs, we could see that literature does not change in the sudden or conjunctural manner which we envisioned at the project's outset. What we refer to as the Underwood hypothesis, a model which is explicitly committed to this understanding of literature only changing over a very long period of time, was therefore validated and remained robust even when methods are used which make the explicit attempt to identify sudden breaks or changes in literary history.

The fourth and final chapter attempted to render our results and a broader CLS discourse coherent within literary-critical historiography, especially a literary-critical historiography of an avowedly Marxian tendency. As this chapter explained, the teleological overtones which are said to attach themselves to any explicitly diachronic method, especially those which make significant use of statistics, have become somewhat unpopular within a literary-critical discourse which is more culturalist than historicist in orientation. The commercialisation of universities and university outputs have only served to render critical approaches oriented around the projection of a social totality increasingly difficult to sustain. As CLS scholars, we are therefore caught between two opposed positions. We could choose an account of modernisation which would represent literary innovation as ascending piecemeal in correlation with processes of social modernisation or industrialisation. Alternatively, we could deny the existence of this objective tendency altogether on the basis that it merely represents an outgrowth of statistical techniques in general, an argument which we might encounter in some attempted rejections of the application of digital tools to the humanities; such criticisms often terminate in a preference for asynchrony.

This thesis has proposed a dialectical means of synthesising these two positions, which is capable of identifying individual examples within the historiography while retaining a sense of an overarching objective tendency which may transform itself further under particular conditions and within specific contexts. This was achieved by identifying sections of individual texts particularly dense with either positively or negatively correlating word types, in a bid to understand how it is that even as words suggestive of narrative description, such as concrete nouns, domestic interiors and parts of the body, increase in their usage over time, while the broader social totality which one would have thought these words are indicative of becomes less coherent or distinct. These seemingly opposed trends can be accounted for by grounding the development of realism over the nineteenth and twentieth century within a Marxist framework which identifies the role of capitalist production in distorting spatial and temporal relations. As capitalism tends towards secular crises, the social totality becomes increasingly difficult to conceive in overall terms and the entities which formerly served to orient the subject no longer cohere to their former extent. It is hoped that this chapter can provide some secure points of departure for locating statistics and computation within the context of a literary-critical historiography which is rich with materials for those seeking to consider some of the problematics associated with the marshalling of objective evidence and large-scale historical narrative within cultural historigraphy in the future.

Significance of Research

To state the central finding of this thesis in the most straightforward terms possible: it can be confirmed that literature does not change in the sudden or conjunctural manner which we envisioned and proposed at this project's outset. Over the approximately one hundred and seventy year period that we have data for, it seems as though literature changes only incrementally. We do not see the supercession of neo-classicism by romanticism in the early nineteenth century, romanticism by realism in the mid nineteenth century and then realism by naturalism or modernism in and around the early twentieth century, contrary to expectations which we may derive from the literary-critical historiography. When we initially implemented our *t*-tests with a view to establishing a benchmark within which we would expect to see some transformation of the literary milieu, indeed the period of time over which literature does seem to undergo a qualitative transformation is so distended or prolonged that there are only a relatively small number of years which we could consider solely from the point of view of their individual contributions to the historical record. In investigating the temporal dynamics of literary change in spite of these obstacles, we were in effect reduced to partitioning our datasets more or less in half and applying a logistic regularisation regression algorithm to the two resultant blocs in order to identify the degree to which our constructed model could tell the two extremes of the dataset apart. As a result, the thesis' central finding might be nuanced in the following terms:

this thesis represents a verification of the Underwood hypothesis and finds that it holds in poetry and drama over the same time period. This latter claim may require some further nuancing as of course the HathiTrust poetry and drama datasets are significantly smaller than the fiction dataset. We might also say that we have added value in the sense that the Underwood hypothesis has remained robust in spite our application of methods which were deployed at least in part with the aim of moving beyond or introducing an alternate perspective to the Underwood model.

There are a number of potential reasons which we could arrive at in speculating as to why the results we obtained were in certain respects, somewhat disappointing. The first and most obvious one, is that we were looking through the wrong end of the telescope in the first instance. Underwood had already proven conclusively in *Distant Horizons* that literature changes slowly and any method which begins with attempting to identify years which are particularly innovative was unlikely to be successful. Secondly, we might propose that the dynamics at work in word frequency and topic model data are to a certain extent incommensurate. It may be the case that distance-based methods applied to topic model frequency data may be more prone to the identification of not just sudden and innovative differences, but a broader spectrum of distinct diachronic behaviours, as the relatively extensive typology of agent behaviour in Barron et al.'s results attest to. However, it should be noted that this thesis has provided concrete information regarding the presence of duplicated texts in the HathiTrust word frequencies data, which has not been dealt with in any comparable depth anywhere else in the literature, as well as some proposed guidelines for dealing with the extensive number of duplicates given the patchy nature of the associated metadata. This thesis has made up some points of overlap in diachronic CLS studies and literary history in such a manner which may provide productive points of departure for future analysts. This was accomplished via a dialectical history of modern literary production steered via a more rigorously Marxian reading of *Capital*.

In this way, this thesis has not necessarily limited itself to the demonstration or replication of a new method but also constitutes a work of literary criticism in its own terms. It is for this reason that individual examples or particular paragraphs were enlisted as opposed to more abstract or broad-ranging instances of distant reading. This thesis has chosen to be more in step with mid-range analyses which are increasingly conducted by CLS scholars working within the field, developing novel ways of rendering the synoptic intelligible (Piper; Underwood, *Distant Horizons: Digital Evidence and Literary Change*). Finally, this thesis has been steered throughout by an objective to render individual, specific and concrete examples of identified changes within germane texts legible within the broader historical narrative it outlines.

Recommendation for Further Research

The first and most pressing necessity for future research conducted within the context of CLS would be the construction of additional literary datasets spread over longer periods of time. The relative lack of word frequency data from before the year 1750 and before, means that no meaningful analysis can be undertaken of how the trends this thesis outlines may relate to literary production in antiquity, the medieval era or the early modern era. In addition to covering a broader historical period of time, it would also be an imperative that these datasets to be more linguistically and geographically diverse than literary datasets or corpora which are available on the internet at the present time. Not only should these proposed corpora provide representative coverage of French, German and Spanish literatures as well as the innumerable regional literatures produced beyond and within the European and American metropoles. It would be also be important that these datasets by agnostic in their structure. Though it has been well-established that word-frequency data can very well represent the actuality of a dataset and literary-change over time and allow analysts to elide copyright restrictions, datasets which provide full texts could allow for the modelling of more involved textual entities in order to compare how sentences, topics or broader word-type entities, such as bigrams, trigrams or topics change over time. That these datasets would need to be in open formats can be taken for granted. Some further boons crucial to future research would include improved metadata which would allow for global implementations of automatic de-duplication tasks and more collaboration between institutions; such that the joining of large literary datasets may become more straightforward. Now that computational methods with a proven track record of functioning in such a way that allows for the diachronic analysis of literature, the future of CLS will lie in the direction of providing the necessary infrastructure to enlarge on analyses.

One futher dynamic which might be proposed is a means of comprehending not just temporal change but interchange. One of the most exciting aspects of Barron et al.'s method is the capacity the method demonstrates for perceiving how influence can mutually enhance two distinct agents within a dataset. Once CLS has developed sufficient datasets and methods to move beyond the ascending or descending line towards more dynamic models which can reproduce the struggle at the crux of literary change, CLS may truly aspire to the culmination of the literary-critical totality of the kind proposed by Casanova in *The World Republic of Letters* (Casanova). Eder has already outlined the potential offered by the modelling of asymmetric relationships in literary corpora (Eder, "Visualization in stylometry: Cluster analysis using networks") and developments underway in the field of network analysis seem to offer some more promising avenues for the future in this regard (Brandes and Erlebach 1; Barthélemy; Brinkmeier and Schank; Fortunato and Hric; Masuda et al.; Zanin et al.).

For all the criticisms which can and should be made of the Marxian approach to the criticism of literature, the material available in Marx's critique of political economy provides a potent means of proceeding, especially given the extent of empirical data, calculus, and algebra Marx employs in support of his more wide-ranging historical theories. It is important to note that literary critics without the benefit of computational methods have anticipated our results regarding the accretion of matter and objects in literary space, while the conditions from which they arise are simultaneously attenuated (Fletcher and Bradbury; Mulhern 778) and this underlines all the more the imperative for analyses of the kind that this thesis has demonstrated the practicability of to become more general within literary criticism as a discipline. There is no reason for a relativist perspective to predominate within literary criticism when historical evidence is available and it is for this reason crucial that a more totalising approach which can accommodate or be adequate to the notion of literary change over time be constructed. It may equally be the case that some of these hypotheses for future work represent a failure to internalise or even properly come to terms with the simple fact that the dynamism of literary history is not quite as borne out on the level of word frequencies to the degree that we might wish. Nevertheless, to paraphrase Jameson, the addition of ‘a third term’ may change everything (Anderson, *The Origins of Postmodernity* 50).

Appendix A: Delta

```
#for the purposes of this example we will consider three
#paragraphs: txt.1, txt.2 and txt.3

txt.1 <- "The wealth of those societies in which the capitalist
mode of production prevails, presents itself as an immense
accumulation of commodities, its unit being a single commodity.
Our investigation must therefore begin with the analysis of a
commodity."

txt.2 <- "The circular movement of capital takes place in three
stages, which, according to the presentation in Volume I, form
the following series: First stage: The capitalist appears as
a buyer on the commodity- and the labour-market; his money is
transformed into commodities, or it goes through the circulation
act M - C. Second Stage: Productive consumption of the purchased
commodities by the capitalist. He acts as a capitalist producer
of commodities; his capital passes through the process of
production. The result is a commodity of more value than that of
the elements entering into its production. Third Stage: The
```

```
capitalist returns to the market as a seller; his commodities
are turned into money; or they pass through the circulation act
C - M."

txt.3 <- "In Book I we analysed the phenomena which constitute the
process of capitalist production as such, as the immediate
productive process, with no regard for any of the secondary
effects of outside influences. But this immediate process of
production does not exhaust the life span of capital. It is
supplemented in the actual world by the process of circulation,
which was the object of study in Book II. In the latter, namely
in Part III, which treated the process of circulation as a medium
for the process of social reproduction, it developed that the
capitalist process of production taken as a whole represents a
synthesis of the processes of production and circulation."

#we combine our paragraphs into a corpus
my.corpus.raw <- list(txt.1, txt.2, txt.3)

#then tokenize the contents of the corpus
my.corpus.clean <- lapply(my.corpus.raw, txt.to.words)

#the five most frequent words in these three samples are 'the',
#'of', 'a', 'as' and 'in'
my.favourite.words <- c("the", "of", "a", "as", "in")

#These are the MFWs out of which we construct our frequency table
freqs <- make.table.of.frequencies(my.corpus.clean,
         my.favourite.words, absent.sensitive = F)


#our frequency table therefore appears as follows
print(freqs)
```

```
##                the        of        a       as       in
## sample_1  7.894737 10.526316 5.263158 2.631579 2.631579
## sample_2 12.295082  4.918033 3.278688 2.459016 1.639344
## sample_3 11.607143 10.714286 2.678571 3.571429 4.464286
##
## (total number of rows/columns:  3/5)
```

```
#we account for the relative frequencies of txt.1 as follows:


#txt.1 is 38 words in length
#the word 'the' occurs three times in txt.1
(3 / 38) * 100
```

```
## [1] 7.894737
```

```
#the word 'of' occurs four times in txt.2
(4 / 38) * 100
```

```
## [1] 10.52632
```

```
#the word 'a' occurs twice in txt.1
(2 / 38) * 100
```

```
## [1] 5.263158
```

```
#we then transform our relative word frequencies into z-scores


scale(freqs)
```

```
##                 the         of          a         as         in
## sample_1 -1.1424440  0.5485791  1.1259325 -0.4272055 -0.1954826
## sample_2  0.7165362 -1.1542303 -0.3411399 -0.7154406 -0.8878245
## sample_3  0.4259078  0.6056512 -0.7847926  1.1426461  1.0833071
##
## (total number of rows/columns:  3/5)
```

```
#then we apply cosine distance to our relative word frequencies

dist.cosine(freqs)
```

```
##              sample_1   sample_2
## sample_2 0.13516351
## sample_3 0.03846297 0.07228584
```

```
#as can be seen here, the distance between text 1 and text 2 is
#calculated as being equal to 0.14.

#cosine distance is calculated as follows

#first we calculate the dot product of text 1 and text 2

dotproduct <- as.numeric(freqs[1,]) %*% as.numeric(freqs[2,])

#then we calculate each vector's magnitude

textonemagnitude <- sqrt(sum(freqs[1,] ^ 2))
texttwomagnitude <- sqrt(sum(freqs[2,] ^ 2))

#multiply them by each other
```

```
magnituderesult <- textonemagnitude * texttwomagnitude


#divide the dot product by magnitude and subtract this result from 1
1 - (dotproduct / magnituderesult)
```

```
##           [,1]
## [1,] 0.1351635
```

Appendix B: Words

Positively Correlating Words (Fiction)

```
##   [1] "about"          "absently"       "accordance"
##   [4] "across"         "advent"         "afore"
##   [7] "afternoon"      "afterward"      "again"
##  [10] "ago"            "ahead"          "almost"
##  [13] "along"          "altogether"     "always"
##  [16] "amount"         "angry"          "announcement"
##  [19] "annoyance"      "annoyed"        "anybody"
##  [22] "anyhow"         "anyone"         "anything"
##  [25] "anywhere"       "apart"          "appreciation"
##  [28] "aristocratic"   "around"         "arranged"
##  [31] "artistic"       "aside"          "ask"
##  [34] "asked"          "available"      "away"
##  [37] "baby"           "back"           "background"
##  [40] "bah"            "bank"           "based"
##  [43] "bearing"        "beautiful"      "bedroom"
##  [46] "behind"         "bells"          "belongings"
##  [49] "below"          "bent"           "beside"
##  [52] "best"           "better"         "between"
##  [55] "bewilderment"     "big"              "birthday"
##  [58] "bit"              "bitter"           "bitterly"
##  [61] "black"            "blanket"          "blinding"
##  [64] "blue"             "boy"              "boys"
##  [67] "breath"           "brief"            "bright"
##  [70] "broad"            "broken"           "burned"
##  [73] "busy"             "cab"              "came"
##  [76] "cared"            "carefully"        "cars"
##  [79] "case"             "catch"            "caught"
##  [82] "chair"            "chance"           "changed"
##  [85] "childlike"        "christmas"        "chuckle"
##  [88] "chuckled"         "church"           "churchyard"
##  [91] "cigar"            "cigars"           "class"
##  [94] "clear"            "clever"           "click"
##  [97] "close"            "closed"           "clutched"
## [100] "clutching"        "cold"             "colour"
## [103] "come"             "comes"            "coming"
## [106] "commenced"        "companionship"    "continuous"
## [109] "control"          "conventional"     "cool"
## [112] "corner"           "cosy"             "course"
## [115] "craft"            "crept"            "crossed"
## [118] "crowd"            "curly"            "curtly"
## [121] "dark"             "dashed"           "day"
## [124] "dazed"            "dead"             "decided"
## [127] "defiant"          "defiantly"        "definite"
## [130] "definitely"       "dense"            "depths"
## [133] "did"              "dim"              "direction"
## [136] "disappeared"      "disapproval"      "distasteful"
## [139] "do"               "does"             "doggedly"
## [142] "doing"            "dollar"           "dollars"
## [145] "done"             "door"             "down"
## [148] "downstairs"       "dreamily"         "dreamy"
## [151] "drew"             "drifted"          "drifting"
## [154] "drive"            "dropped"          "dull"
## [157] "dusty"            "earnest"          "ears"
## [160] "edge"             "else"             "empty"
## [163] "enough"           "entire"           "everybody"
## [166] "everyone"         "everything"       "everywhere"
## [169] "evidently"        "excitedly"        "excitement"
## [172] "expression"       "eyes"             "face"
## [175] "faces"            "fact"             "fairly"
## [178] "familiar"         "fast"             "feel"
## [181] "feeling"          "feet"             "fellow"
## [184] "felt"             "finally"          "financial"
## [187] "finger"           "fingers"          "first"
## [190] "fixedly"          "flashed"          "flickering"
## [193] "floor"            "flush"            "fog"
## [196] "foolish"          "forever"          "forgotten"
## [199] "fresh"            "front"            "full"
## [202] "funny"            "furtively"        "gathered"
## [205] "gathering"        "gaze"             "gently"
## [208] "get"              "getting"          "girl"
## [211] "girlhood"         "girlish"          "girls"
## [214] "glad"             "glance"           "glanced"
## [217] "glancing"         "go"               "goes"
## [220] "goin'"            "going"            "gone"
## [223] "got"              "grass"            "gravely"
## [226] "green"            "greeting"         "grimly"
## [229] "growing"          "guests"           "hair"
## [232] "half"             "handed"           "hands"
## [235] "handwriting"      "hard"             "hat"
## [238] "hauled"           "head"             "heavily"
## [241] "heavy"            "held"             "help"
## [244] "helped"           "here"             "holding"
## [247] "holiday"          "home"             "homes"
## [250] "hopelessly"       "horrified"        "horses"
## [253] "house"            "household"        "hurriedly"
## [256] "husky"            "i'll"             "ignore"
## [259] "ignored"          "ignoring"         "imaginative"
## [262] "indoors"          "inquiringly"      "inside"
## [265] "instinct"         "intense"          "jerked"
## [268] "jim"              "just"             "keen"
## [271] "keep"             "keeping"          "kept"
## [274] "kin"              "kiss"             "knew"
## [277] "know"             "lane"             "later"
## [280] "laugh"            "laughed"          "laughing"
## [283] "lazily"           "leaned"           "leaning"
## [286] "lifted"           "light"            "lighted"
## [289] "like"             "liked"            "likes"
## [292] "line"             "lips"             "listen"
## [295] "listening"        "little"           "locality"
## [298] "log"              "logs"             "lonely"
## [301] "look"             "looked"           "looking"
## [304] "loomed"           "loving"           "low"
## [307] "lower"            "lunch"            "luncheon"
## [310] "lying"            "mainly"           "managed"
## [313] "matter"           "maybe"            "memories"
## [316] "minutes"          "mood"             "moonlight"
## [319] "moustache"        "move"             "moved"
## [322] "movement"         "moving"           "murmured"
## [325] "muttered"         "n't"              "narrow"
## [328] "nearing"          "nearly"           "neck"
## [331] "need"             "needed"           "nervous"
## [334] "nervously"        "nervousness"      "new"
## [337] "nice"             "night"            "nightmare"
## [340] "nodded"           "noiselessly"      "nonsense"
## [343] "nose"             "noticed"          "nowadays"
## [346] "occupant"         "occupants"        "off"
## [349] "old"              "older"            "once"
## [352] "open"             "opened"           "out"
## [355] "outbreak"         "outburst"         "outside"
## [358] "over"             "overcoat"         "overhead"
## [361] "package"          "pale"             "passionate"
## [364] "past"             "pathway"          "peered"
## [367] "peering"          "penniless"        "persistent"
## [370] "persistently"     "phase"            "piano"
## [373] "pleasant"         "policeman"        "position"
## [376] "practical"        "pretty"           "programme"
## [379] "pull"             "pushed"           "quaint"
## [382] "queer"            "question"         "quick"
## [385] "quickly"          "quiet"            "quietly"
## [388] "quite"            "railroad"         "railway"
## [391] "rapidly"          "rate"             "reached"
## [394] "reaction"         "reckless"         "reckon"
## [397] "recognised"       "red"              "remark"
## [400] "remarked"         "remember"         "remembered"
## [403] "reminiscences"    "reply"            "respond"
## [406] "responded"        "responsibilities" "resting"
## [409] "restlessly"       "result"           "revolver"
## [412] "ride"             "riding"           "rifle"
## [415] "right"            "ripple"           "roadside"
## [418] "rolled"           "room"             "rooms"
## [421] "rose"             "rough"            "round"
## [424] "rush"             "sad"              "sadly"
## [427] "said"             "sat"              "save"
## [430] "say"              "school"           "see"
## [433] "seemed"           "seen"             "sewing"
## [436] "shadow"           "shake"            "shaking"
## [439] "sharp"            "sharply"          "shawl"
## [442] "shining"          "shook"            "shot"
## [445] "shoulder"         "shoulders"        "shouted"
## [448] "shouting"         "show"             "shyly"
## [451] "sick"             "side"             "sign"
## [454] "simply"           "simultaneously"   "singing"
## [457] "sitting"          "sky"              "sleep"
## [460] "sleeping"         "slight"           "slightly"
## [463] "slopes"           "slow"             "slowly"
## [466] "small"            "smile"            "smiled"
## [469] "smoking"          "snow"             "soft"
## [472] "softly"           "somebody"         "somehow"
## [475] "something"        "somewhat"         "somewhere"
## [478] "soothingly"       "sort"             "sound"
## [481] "southern"         "speak"            "speaking"
## [484] "special"          "spoke"            "spoken"
## [487] "sprang"           "stalwart"         "stand"
## [490] "standing"         "start"            "started"
## [493] "startled"         "startling"        "stealthily"
## [496] "stealthy"         "steamer"          "step"
## [499] "stepped"          "stern"            "stiffly"
## [502] "still"            "stone"            "stood"
## [505] "stop"             "store"            "stove"
## [508] "straight"         "straightforward"  "strange"
## [511] "strangely"        "street"           "streets"
## [514] "strong"           "struggle"         "suddenly"
## [517] "suggested"        "suggestive"       "summer"
## [520] "sunlight"         "sunny"            "sunset"
## [523] "sunshine"         "suppose"          "sure"
## [526] "surged"           "surroundings"     "swaying"
## [529] "swept"            "table"            "tact"
## [532] "talk"             "talking"          "tall"
## [535] "tea"              "teacher"          "team"
## [538] "teeth"            "telegram"         "tell"
## [541] "temporarily"      "thank"            "then"
## [544] "there"            "thin"             "things"
## [547] "think"            "thinking"         "thoroughly"
## [550] "thoughtfully"     "through"          "tightened"
## [553] "tightly"          "tiny"             "tired"
## [556] "tone"             "tones"            "touch"
## [559] "touched"          "toward"           "trail"
## [562] "tramp"            "tried"            "trouble"
## [565] "troubled"         "truthful"         "try"
## [568] "trying"           "turned"           "turning"
## [571] "understand"       "undertone"        "uneasily"
## [574] "unfamiliar"       "unmistakable"     "unselfish"
## [577] "until"            "up"               "upstairs"
## [580] "upturned"         "vague"            "vaguely"
## [583] "veritable"        "vividly"          "voice"
## [586] "voices"           "waiting"          "walked"
## [589] "want"             "watch"            "watched"
## [592] "watching"         "way"              "wearily"
## [595] "week"             "well"             "went"
## [598] "whispered"        "white"            "why"
## [601] "wide"             "winced"           "window"
## [604] "windows"          "womanly"          "wonder"
## [607] "wondering"        "words"            "wore"
## [610] "work"             "worry"            "wrong"
## [613] "yearning"         "years"            "yelled"
## [616] "yellow"           "yes"              "you're"
```

Negatively Correlating Words (Fiction)

```
##   [1] "abilities"       "accident"        "accompanied"     "accompany"
##   [5] "accomplishments" "account"         "acknowledge"     "acknowledged"
##   [9] "acquainted"      "acquire"         "acquired"        "actions"
##  [13] "add"             "address"         "addresses"       "adieu"
##  [17] "admiration"      "advantage"       "advantages"      "advice"
##  [21] "affected"        "affecting"       "affection"       "affections"
##   [25] "afflicted"     "affliction"    "afforded"      "agitated"
##   [29] "agitation"     "agonies"       "agreeable"     "agreed"
##   [33] "alarmed"       "alarming"      "alas"          "alliance"
##   [37] "amiable"       "amply"         "amusements"    "animated"
##   [41] "any"           "apartment"     "apology"       "appear"
##   [45] "appeared"      "appears"       "applied"       "apply"
##   [49] "apprehension"  "apprehensions" "apprehensive"  "approbation"
##   [53] "approve"       "approved"      "ardent"        "ardour"
##   [57] "arguments"     "arid"          "arising"       "artful"
##   [61] "arts"          "assist"        "assistance"    "assisted"
##   [65] "assurances"    "assured"       "assuring"      "attachment"
##   [69] "attempted"     "attend"        "attended"      "attending"
##   [73] "attention"     "attentive"     "aversion"      "avoid"
##   [77] "avowed"        "banished"      "barbarous"     "begged"
##   [81] "behaviour"     "beheld"        "being"         "beings"
##   [85] "beloved"       "benefactor"    "benevolence"   "benevolent"
##   [89] "bestow"        "bestowed"      "birth"         "blessings"
##   [93] "bosom"         "bounds"        "bounty"        "bred"
##   [97] "candour"       "capable"       "cause"         "caution"
##  [101] "cease"         "censure"       "characters"    "charms"
##  [105] "cheerfulness"  "choice"        "circumstance"  "civility"
##  [109] "commands"      "communicate"   "communicated"  "company"
##  [113] "compassion"    "complain"      "complaints"    "compliments"
##  [117] "comply"        "compose"       "composed"      "conceal"
##  [121] "concealed"     "concealing"    "conceive"      "conceived"
##  [125] "concern"       "concerns"      "conclude"      "concluded"
##  [129] "concluding"    "condemn"       "condescension" "conduct"
##  [133] "conducted"     "confine"       "confined"      "confinement"
## [137] "confirm"         "confirmed"       "confusion"       "connexions"
## [141] "consent"         "consented"       "consequence"     "consequences"
## [145] "consider"        "considered"      "considering"     "consolation"
## [149] "contain"         "contempt"        "contented"       "continual"
## [153] "continue"        "continuing"      "contrary"        "contribute"
## [157] "contributed"     "convenience"     "conveyed"        "convince"
## [161] "convinced"       "countenance"     "country"         "cruel"
## [165] "cruelty"         "cultivate"       "curiosity"       "danger"
## [169] "dangers"         "deceived"        "decency"         "declaration"
## [173] "declare"         "degree"          "delicacy"        "deliver"
## [177] "delivered"       "denied"          "departure"       "depend"
## [181] "depends"         "deprive"         "deprived"        "describe"
## [185] "deserve"         "deserved"        "deserves"        "deserving"
## [189] "design"          "designs"         "desired"         "desiring"
## [193] "desirous"        "despair"         "destitute"       "destroy"
## [197] "destroyed"       "detain"          "detained"        "determine"
## [201] "determined"      "dictated"        "directed"        "disappointment"
## [205] "disclose"        "discourse"       "discover"        "discovered"
## [209] "discovering"     "discovery"       "discretion"      "disinterested"
## [213] "disorder"        "disordered"      "dispatched"      "displeased"
## [217] "displeasure"     "dispose"         "disposition"     "dispute"
## [221] "dissipated"      "distinction"     "distraction"     "distress"
## [225] "distressed"      "divert"          "doubted"         "dreaded"
## [229] "dreadful"        "ease"            "education"       "effects"
## [233] "elegance"        "elegant"         "eloquence"       "embrace"
## [237] "embraced"        "embracing"       "emotions"        "employed"
## [241] "employment"      "encourage"       "encouraged"      "endeavour"
## [245] "endeavoured"     "endeavouring"    "endeavours"      "enemies"
## [249] "engage"            "engagements"     "engaging"        "enjoy"
## [253] "enjoyed"           "entertain"       "entertained"     "entertainment"
## [257] "entreat"           "entreated"       "envy"            "equal"
## [261] "equally"           "error"           "errors"          "esteem"
## [265] "esteemed"          "event"           "every"           "evils"
## [269] "exalted"           "examine"         "example"         "excess"
## [273] "excite"            "execute"         "executed"        "exert"
## [277] "exerted"           "exertion"        "expectation"     "expectations"
## [281] "expense"           "experienced"     "expired"         "expose"
## [285] "exposed"           "express"         "expressing"      "expressions"
## [289] "expressive"        "extend"          "extraordinary"   "extravagance"
## [293] "extreme"           "extremely"       "fail"            "fame"
## [297] "fatal"             "fatigue"         "favour"          "favourable"
## [301] "favours"           "fears"           "fee"             "felicity"
## [305] "female"            "fidelity"        "filial"          "flatter"
## [309] "flattered"         "flattering"      "flew"            "flow"
## [313] "folly"             "forbid"          "forcibly"        "formed"
## [317] "former"            "formerly"        "fortitude"       "fortune"
## [321] "foundation"        "founded"         "frequent"        "frequently"
## [325] "friend"            "friendship"      "gaiety"          "generosity"
## [329] "generous"          "goodness"        "governed"        "graces"
## [333] "granted"           "gratification"   "gratify"         "gratitude"
## [337] "greater"           "greatest"        "grief"           "guilty"
## [341] "habitation"        "happiness"       "hasten"          "having"
## [345] "health"            "hearing"         "heaven"          "hereafter"
## [349] "highly"            "hitherto"        "honour"          "honours"
## [353] "hopes"             "horror"          "horrors"         "however"
## [357] "humanity"          "humour"          "ideas"           "ignorant"
## [361] "ill"             "imagination"     "imagined"        "immediate"
## [365] "immediately"     "impart"          "impatience"      "impatient"
## [369] "impetuous"       "impose"          "imposed"         "improper"
## [373] "improve"         "improved"        "incapable"       "inclination"
## [377] "inclinations"    "increase"        "increased"       "indebted"
## [381] "induce"          "induced"         "indulge"         "indulged"
## [385] "indulgence"      "infancy"         "inferior"        "infinitely"
## [389] "inform"          "informed"        "informing"       "ingratitude"
## [393] "inhabitants"     "injured"         "innocence"       "innocent"
## [397] "inquire"         "insensible"      "inspire"         "inspired"
## [401] "instantly"       "instruct"        "integrity"       "intelligence"
## [405] "intended"        "intention"       "intentions"      "interrupt"
## [409] "introduced"      "journey"         "joy"             "judgement"
## [413] "justice"         "justly"          "lament"          "lamented"
## [417] "language"        "lasting"         "lately"          "leave"
## [421] "length"          "lest"            "liberty"         "lively"
## [425] "lodged"          "loss"            "lustre"          "madam"
## [429] "mankind"         "manner"          "manners"         "marks"
## [433] "may"             "means"           "measures"        "meditated"
## [437] "melancholy"      "mention"         "mentioned"       "merit"
## [441] "merits"          "method"          "mind"            "minds"
## [445] "miseries"        "misery"          "misfortune"      "misfortunes"
## [449] "modesty"         "mortification"   "mortified"       "most"
## [453] "motive"          "motives"         "mould"           "mutual"
## [457] "my"              "myself"          "nature"          "necessity"
## [461] "neglect"         "neither"         "nor"             "notwithstanding"
## [465] "object"          "objection"       "objects"         "obligation"
## [469] "obligations"     "oblige"          "obliged"         "obliging"
## [473] "observation"     "observations"    "observe"         "observing"
## [477] "obtain"          "obtained"        "occasion"        "occasioned"
## [481] "occasions"       "offended"        "offers"          "opinion"
## [485] "opportunity"     "oppose"          "opposed"         "ordered"
## [489] "ornament"        "overwhelmed"     "owe"             "owed"
## [493] "pains"           "pangs"           "parent"          "part"
## [497] "partake"         "partial"         "particular"      "particularly"
## [501] "particulars"     "passion"         "passions"        "paternal"
## [505] "peace"           "perceive"        "perceived"       "perceiving"
## [509] "perform"         "period"          "permit"          "person"
## [513] "persuade"        "persuaded"       "persuasion"      "piqued"
## [517] "pleased"         "pleasing"        "pleasure"        "pleasures"
## [521] "politeness"      "power"           "preceding"       "preference"
## [525] "prejudices"      "preparing"       "present"         "preserve"
## [529] "preserved"       "presumption"     "prevail"         "prevailed"
## [533] "prevent"         "prevented"       "principles"      "probability"
## [537] "proceed"         "procure"         "procured"        "produce"
## [541] "promised"        "pronounce"       "proof"           "proofs"
## [545] "proper"          "proportion"      "proposal"        "propriety"
## [549] "protection"      "prove"           "provided"        "providence"
## [553] "prudence"        "prudent"         "pursue"          "pursuing"
## [557] "qualities"       "quit"            "quitted"         "quitting"
## [561] "rage"            "rank"            "rational"        "readily"
## [565] "reason"          "reasons"         "receive"         "received"
## [569] "receiving"       "reception"       "recollection"    "recommend"
## [573] "recommended"     "reconcile"       "reconciled"      "recourse"
## [577] "recover"         "recovered"       "recovering"      "recovery"
## [581] "redoubled"       "reduced"         "reflect"         "reflecting"
## [585] "reflection"    "reflections"   "regard"        "regret"
## [589] "rejected"      "rejoice"       "rejoiced"      "relate"
## [593] "related"       "relieve"       "reluctance"    "remainder"
## [597] "remaining"     "remains"       "remove"        "render"
## [601] "rendered"      "rendering"     "renew"         "repair"
## [605] "repeat"        "repeatedly"    "repent"        "repentance"
## [609] "repose"        "reposed"       "reproach"      "reproaches"
## [613] "request"       "requisite"     "resentment"    "reserve"
## [617] "residence"     "resign"        "resignation"   "resolution"
## [621] "resolved"      "resource"      "respect"       "restore"
## [625] "restored"      "restraint"     "retain"        "retire"
## [629] "retired"       "retirement"    "retiring"      "retreat"
## [633] "return"        "returning"     "revive"        "reward"
## [637] "rewarded"      "riches"        "ridicule"      "rival"
## [641] "satisfaction"  "scruple"       "secure"        "security"
## [645] "sensations"    "sensibility"   "sensible"      "sentiment"
## [649] "sentiments"    "separation"    "severe"        "severely"
## [653] "severity"      "sex"           "simplicity"    "sincere"
## [657] "sincerely"     "sincerity"     "situation"     "soften"
## [661] "solicitude"    "sooner"        "sorrows"       "spirits"
## [665] "subjects"      "submit"        "suffer"        "suffered"
## [669] "sufferings"    "sufficient"    "sunk"          "superior"
## [673] "supplied"      "support"       "supported"     "supposition"
## [677] "suppress"      "suspicions"    "talents"       "taste"
## [681] "taught"        "tear"          "tedious"       "temper"
## [685] "tender"        "tenderest"     "tenderness"    "terms"
## [689] "terrified"     "therefore"     "this"          "thither"
## [693] "thus"          "ties"          "till"          "title"
## [697] "tolerable"       "totally"         "tranquillity"    "transport"
## [701] "transported"     "treated"         "trifling"        "unable"
## [705] "unbounded"       "uncertainty"     "uncommon"        "understanding"
## [709] "uneasiness"      "unequal"         "unfortunate"     "ungrateful"
## [713] "unhappy"         "united"          "universally"     "unwilling"
## [717] "unworthy"        "utmost"          "vain"            "value"
## [721] "vanity"          "variety"         "vexation"        "vice"
## [725] "vices"           "view"            "violence"        "violent"
## [729] "virtue"          "virtues"         "virtuous"        "visits"
## [733] "vivacity"        "warmest"         "weakness"        "whence"
## [737] "whither"         "whom"            "wishes"          "wit"
## [741] "without"         "worthy"          "wretch"          "wretchedness"
## [745] "wretches"        "zeal"
```

Positively Correlating Words (Poetry)

```
##   [1] "about"      "above"      "across"     "after"      "again"
##   [6] "aglow"      "ago"        "all"        "am"         "among"
##  [11] "and"        "angel"      "angels"     "another"    "anything"
##  [16] "apart"      "are"        "as"         "asked"      "asleep"
##  [21] "autumn"     "away"       "back"       "bars"       "be"
##  [26] "beautiful"  "before"     "beside"     "beyond"     "bird"
##  [31] "birds"      "bitter"     "blue"       "break"      "bring"
##  [36] "brings"     "broken"     "brought"    "burden"     "calling"
##  [41] "came"       "cannot"     "cared"      "child"      "children"
##  [46] "clear"      "come"       "comes"      "cometh"     "coming"
##  [51] "crept"      "darling"    "dawn"       "day"        "days"
##  [56] "dead"       "dear"       "dim"        "do"         "done"
##  [61] "door"       "down"       "dream"      "dreamed"    "dreaming"
##  [66] "dreams"     "dreamy"     "dropped"    "earnest"    "earth"
##  [71] "ever"       "everything" "everywhere" "evil"       "eyes"
##  [76] "face"       "faces"      "faith"      "falling"    "far"
##  [81] "fashioned"  "feet"       "folded"     "for"        "free"
##  [86] "fresh"      "gather"     "gathered"   "gift"       "glad"
##  [91] "gladness"   "go"         "god"        "gold"       "golden"
##  [96] "gone"       "grand"      "grass"      "grasses"    "greed"
## [101] "grew"       "grey"       "grow"       "growing"    "grown"
## [106] "grows"      "hair"       "hands"      "have"       "hear"
## [111] "heart"      "hearts"     "held"       "help"       "hidden"
## [116] "higher"     "hold"       "home"       "into"       "is"
## [121] "it"         "its"        "jewelled"   "keep"       "kissed"
## [126] "knew"       "know"       "land"       "last"       "life"
## [131] "light"      "lilies"     "lips"       "lit"        "little"
## [136] "living"     "longing"    "looked"     "looking"    "love"
## [141] "loved"      "loving"     "low"        "lying"      "made"
## [146] "maiden"     "manhood"    "may"        "me"         "memories"
## [151] "men"        "mine"       "mission"    "mist"       "morning"
## [156] "mother"     "music"      "mystery"    "nearer"     "need"
## [161] "needs"      "nest"       "never"      "night"      "not"
## [166] "old"        "olden"      "one"        "ones"       "only"
## [171] "our"        "ours"       "out"        "over"       "passed"
## [176] "past"       "pathway"    "perfect"    "place"      "pleasant"
## [181] "prayer"     "precious"   "priceless"  "quiet"      "rain"
## [186] "rest"       "right"      "ringing"    "ripple"     "ripples"
## [191] "river"      "room"       "said"       "sat"        "say"
## [196] "sea"        "see"        "seems"      "set"        "shadow"
## [201] "shadows"    "shining"    "sing"       "singing"    "sky"
## [206] "sleep"      "sleeping"   "slowly"     "snow"       "so"
## [211] "softly"     "something"  "songs"      "speak"      "stand"
## [216] "standing"   "star"       "stars"      "stayed"     "stilled"
## [221] "stirred"    "story"      "strange"    "strife"     "strong"
## [226] "suddenly"   "summer"     "sun"        "sunlight"   "sunny"
## [231] "sunrise"    "sunset"     "sunshine"   "surely"     "swaying"
## [236] "sweet"      "sweeter"    "swing"      "take"       "talked"
## [241] "tall"       "tell"       "that"       "them"       "there"
## [246] "things"     "through"    "time"       "tiny"       "together"
## [251] "toward"     "tree"       "trees"      "true"       "trust"
## [256] "turned"     "until"      "up"         "upon"       "upward"
## [261] "us"         "violets"    "vision"     "voices"     "wait"
## [266] "waiting"    "walked"     "wandered"   "way"        "we"
## [271] "weary"      "weird"      "went"       "were"       "west"
## [276] "whisper"    "whispered"  "white"      "will"       "win"
## [281] "within"     "won"        "word"       "words"      "work"
## [286] "would"      "wrong"      "year"       "yearning"   "years"
## [291] "you"
```

Negatively Correlating Words (Poetry)

```
##   [1] "adorn"     "age"       "aid"       "ambition"  "anxious"   "appear"
##   [7] "appears"   "approach"  "arm"       "arms"      "array"     "arts"
##  [13] "attend"    "beam"      "betray"    "bids"      "blaze"     "boast"
##  [19] "bosom"     "bow"       "breast"    "cause"     "charm"     "charms"
##  [25] "command"   "conscious" "convey"    "descend"   "design"    "destroy"
##  [31] "dire"      "display"   "doom"      "dread"     "dreadful"  "each"
##  [37] "ease"      "employ"    "envy"      "equal"     "exulting"  "eye"
##  [43] "fame"      "fancy"     "fatal"     "fate"      "fix"       "flies"
##  [49] "flow"      "form"      "former"    "gale"      "grateful"  "groan"    
##  [55] "grove"     "groves"    "hence"     "his"       "impatient" "join"     
##  [61] "labours"   "loud"      "lustre"    "lyre"      "mark"      "mild"     
##  [67] "mind"      "mourn"     "muse"      "native"    "nature"    "nor"      
##  [73] "oft"       "parent"    "passions"  "pleasure"  "pour"      "prey"     
##  [79] "pride"     "proud"     "pursue"    "rage"      "rear"      "reason"   
##  [85] "reign"     "resign"    "return"    "rude"      "ruin"      "sacred"   
##  [91] "scene"     "severe"    "shades"    "sigh"      "spread"    "spreads"  
##  [97] "stray"     "sunk"      "survey"    "sway"      "taste"     "tear"     
## [103] "thus"      "train"     "trembling" "turn"      "tyrant"    "vain"     
## [109] "various"   "vice"      "view"      "views"     "virtue"    "virtuous" 
## [115] "woes"      "wretch"    "youth"     "zeal"
```

Positively Correlating Words (Drama)

```
##   [1] "about"      "across"     "afraid"     "after"      "afternoon" 
##   [6] "ago"        "always"     "any"        "anyone"     "anything"  
##  [11] "anyway"     "around"     "asked"      "at"         "awfully"   
##  [16] "back"       "because"    "been"       "beginning"  "begins"    
##  [21] "behind"     "big"        "business"   "carries"    "catches"   
##  [26] "chair"      "coming"     "course"     "curtain"    "did"       
##  [31] "different"  "do"         "does"       "doing"      "door"      
##  [36] "doorway"    "down"       "enters"     "everyone"   "everything"
##  [41] "excitedly"  "exits"      "facing"     "finally"    "fireplace" 
##  [46] "get"        "gets"       "getting"    "girl"       "glad"      
##  [51] "glancing"   "go"         "goes"       "going"      "got"       
##  [56] "hands"      "he"         "hello"      "hurry"      "indicating"
##  [61] "intently"   "interested" "into"       "it"         "jumps"     
##  [66] "just"       "know"       "later"      "laughing"   "listens"
##  [71] "little"     "looking"    "looks"      "loudly"     "lunch"
##  [76] "matter"     "mean"       "minute"     "money"      "nervously"
##  [81] "off"        "only"       "out"        "outside"    "over"
##  [86] "paces"      "piano"      "picking"    "puts"       "quickly"
##  [91] "quietly"    "reaches"    "realise"    "really"     "remember"
##  [96] "right"      "rises"      "room"       "saying"     "sees"
## [101] "shakes"     "she"        "shoulder"   "simply"     "sit"
## [106] "sits"       "slightly"   "softly"     "someone"    "something"
## [111] "sorry"      "speaking"   "stage"      "standing"   "stands"
## [116] "starts"     "stops"      "suppose"    "table"      "takes"
## [121] "talking"    "tea"        "think"      "thinking"   "today"
## [126] "together"   "told"       "tonight"    "toward"     "trying"
## [131] "turning"    "turns"      "understand" "until"      "up"
## [136] "upstairs"   "want"       "wanted"     "was"        "way"
## [141] "window"     "working"    "yes"        "you"
```

Negatively Correlating Words (Drama)

```
##  [1] "bless"  "brave"  "by"     "cause"  "death"  "earth"  "enter"  "eye"
##  [9] "fear"   "heart"  "heaven" "hour"   "its"    "life"   "may"    "my"
## [17] "nature" "not"    "power"  "shall"  "should" "soul"   "spare"  "speak"
## [25] "such"   "their"  "these"  "this"   "though" "thus"   "upon"   "which"
## [33] "whom"   "whose"  "yet"
```

Appendix C: Creating Frequency Tables

```r
#first we read in our .csv
fiction_metadata <- read.csv("fiction_metadata_norm.csv",
                             stringsAsFactors = F)

#make it a tibble
fiction_metadata <- as_tibble(fiction_metadata)

#drop the number column
fiction_metadata$X <- NULL

#then we read in our roman numerals,
#correction rules and variant spellings
romans <- read.csv("romannumerals.txt",
                   stringsAsFactors = F)
corrections <- read.csv("CorrectionRules.txt",
                   stringsAsFactors = F, sep = "\t", header = F)
variants <- read.csv("VariantSpellings.txt",
                   stringsAsFactors = F, sep = "\t", header = F)

#we then create freqs which will hold all
#the data to be generated in the loop below
freqs <- list(length = length(fiction_metadata$htid))

setwd("fiction")

#for as long as i is less than the length of fictionone
for(i in 1:length(fiction_metadata$htid)) {
  #if ab id string in fiction_metadata, with .tsv
  #at the end is in the directory read the table
  #into a new variable called data
  data <- read.table(fiction_metadata$htid[i],
                     quote = "", stringsAsFactors = FALSE,
                     header = FALSE, fill = TRUE)
  #next, we replace the frequencies appearing in the
  #second column of data with their relative frequencies,
  #expressed as a percentage
  data[,2] <- (data[,2] / sum(data[,2], na.rm = TRUE)) * 100
  #if there are more than or equal to 8695 words in the corpus,
  #take the first 8695 words, turn them into a dataframe
  #and make it the ith element in our freqs list
  freqs[[i]] <- as.data.frame(cbind(data[1:8695,],
                                    fiction_metadata$htid[i]))
}

#we then bind freqs by the ID variable
```

```
freqs <- bind_rows(freqs, .id = "V1")

#change the colnames to facilitate
#its transformation into a tibble
colnames(freqs) <- c("num", "Word", "Frequency", "htid")

#create freqs as a tibble of itself
freqs <- as_tibble(freqs)

freqs$num <- NULL

#remove these variables from our workspace
rm(data, i)

#then we remove all the details we don't want
freqs <- freqs %>% filter(!Word %in% romans[,1])
freqs$Word <- mapvalues(freqs$Word,
              corrections$V1, corrections$V2)
freqs$Word <- mapvalues(freqs$Word,
              variants$V1, variants$V2)

#and then join it with our metadata
freqs <- full_join(freqs, fiction_metadata)

#drop the htid vector as
#we no longer need these
freqs$htid <- NULL
freqs$author <- NULL
```

```
freqs$num <- NULL

#remove these tibbles which we don't need anymore
rm(romans, corrections, variants, fiction_metadata)

#extract the 9019 most frequent words, and we will accomplish this
#by removing words which are not really words, or words we are not
#interested in from the topwords vector
topwords <- sort(table(freqs$Word), decreasing = T)[1:9010]
topwords <- as.data.frame(topwords)
topwords <- as.character(topwords$Var1)
topwords <- sort(topwords)

#remove positions from our workspace as we no longer need it
rm(positions)

freqs <- freqs %>% filter(Word %in% topwords)

detach(package:plyr)

#average by volumes
freqs <- freqs %>%
  group_by(title, enumcron, date, Word) %>%
  summarise(Frequency = mean(Frequency))

#average by title
freqs <- freqs %>%
  group_by(title, date, Word) %>%
  summarise(Frequency = mean(Frequency))

#and average by year
freqs <- freqs %>%
  group_by(date, Word) %>%
  summarise(Frequency = mean(Frequency))

#create data which will hold our freqs
data <- freqs

#then we need to identify which years
#don't have the full 8710 topwords
lowyears <- table(data$date)
lowyears <- as.data.frame(lowyears)
lowyears <- as_tibble(lowyears)

#we then filter out every year which
#doesn't appear in the tibble 8710 times
lowyears <- lowyears %>% filter(lowyears$Freq < length(topwords))

#replace lowyears with its first index
lowyears <- lowyears[[1]]

#remake lowyears as a character
lowyears <- as.character(lowyears)

#and then a numerical vector
lowyears <- as.numeric(lowyears)
```

```r
#we then create a loop which will run as many
#times as we have years missing word data
for(i in 1:length(lowyears)) {

  #first we create a placeholder variable
  #which will hold the data at the specified year
  attempt <- data %>% filter(date == lowyears[i])
  #next we identify the words which are missing from our data
  word <- setdiff(topwords, attempt$Word)
  #we create a vector of zeros corresponding
  #in length to the number of words missing
  frequency <- rep(0, length(word))
  #and a vector of the number of years
  year <- rep(lowyears[i], length(word))
  #we join these together
  columns <- cbind(year, word, frequency)
  #and turn the result into a tibble
  columns <- as_tibble(columns)
  #replace the columns we need to
  #with the data type version we need to
  columns$year <- as.integer(columns$year)
  columns$frequency <- as.double(columns$frequency)
  #change the column names
  colnames(columns) <- c("date", "Word", "Frequency")
  #and then finally, join this second
  ¢placeholder with our original data
  data <- full_join(data, columns)
}

#the final thing to do is to remove
#all NAs by making them equal to zero
nalocations <- which(is.na(data$Frequency))

#replace all the nas with zeroes
data$Frequency[nalocations] <- 0

#we spread the data according to word and year
data <- data %>% spread(Word, Frequency, sep = "")

#scale the data
for(i in 2:dim(data)[2]) {
  data[,i] <- scale(data[,i])
}

#save data
saveRDS(data, file = "fiction_frequencies.rds")
```

Appendix D: T-Tests

```
#readRDS file
fictiondata <- readRDS("fiction_frequencies.rds")

#create fictionyearvector which is going to
#hold all our dates so we can call on them easily
fictionyearvector <- fictiondata$date

fictiondata$date <- NULL

#then we remove the first four characters from
#the colnames of fictiondata so that the
#colnames can be accurate
for(i in 1:length(colnames(fictiondata))) {
  colnames(fictiondata)[i] <- substr(colnames(fictiondata)[i],
  5, nchar(colnames(fictiondata)[i]))
}

fictiondata <- as.matrix(fictiondata)
```

```
#convert fictiondata into a fictiondataframe
#so we can run a distance measurement
fictionn <- as.matrix(dist.cosine(fictiondata))

#create our dummy variable test
fictiontest <- 0

#we want to create a fictiondataframe which is
#every scale from 1:85, and gets added 1 to every
#time a t-test at that scale is significantly different
#and then slots the magnitude of that difference onto
#the end of the row in a continuous vector
fictionfreqs <- 0

#for as long as i is less than the size of our
#distance table
for(i in 1:dim(fictionn)[1]) {
  #we create the variable row, which is equal
  #to the distances obtained from our table
  #where row position = i
  row <- as.numeric(fictionn[i,])
   #if i is less than 11 or more than the size
   #of the distance table divided by two, i.e.
   #if there are less than ten samples on either
   #side of the zero to analyse in our t-tests
  if(i < 11 || i > dim(fictionn)[1]-10) {
   #print no fictiondata
   #if i is less than the size of the distance
   #table divided by two, i.e. if we are still
   #before the halfway point in our distance table
  } else if(i < dim(fictionn)[1] / 2) {
    #one is a new vector equal to the first i
    #digits from our sitance table
    one <- row[1:i-1]
    #two is a corresponding number of figures
    #from the other side of the zero
    two <- row[i+1:length(one)]
    #we remove all the na's from 2
    two <- two[!is.na(two)]
    #now because we want to run a t-test on every
    #possible scale of years, we have to set up a
    #while loop to steadily erode the one and two
    #vectors from either side
    while(length(one) >= 10){
      #run our t-test
      test <- t.test(one, two)
      #we create the variable newrow
      fictionfreqs <- rbind(fictionfreqs,
      c(length(one), test$p.value,
      as.numeric(test$estimate[1]) -
      as.numeric(test$estimate)[2]))
      #we want to remove the first element from one
      one <- one[2:length(one)]
      #and the last element from two
      two <- two[1:length(two) - 1]
      #if the result of our t-test is significant
    }
    #if i is more than the size of
    #the distance table divided by two
  } else {
    #two is these figures corresponding to these indexes
    two <- row[i+1:dim(fictionn)[1]]
    #with the Nas removed
    two <- two[!is.na(two)]
    #we then reverse our row vector
    row <- rev(row)
    #find where zero is in the vector
    position <- match(0, row)
    #and extract the corresponding
    #number of figures from the other side
    one <- row[position + 1:length(two)]
    while(length(one) >= 10){
      test <- t.test(one, two)
      #we create the variable newrow
      fictionfreqs <- rbind(fictionfreqs, c(length(one),
      test$p.value, as.numeric(test$estimate[1]) -
      as.numeric(test$estimate)[2]))
      #we want to remove the last element from one
      one <- one[1:length(one) - 1]
      #and the last element from two
      two <- two[1:length(two) - 1]
    }
  }
}
```

```
#first we remove the first row of fictionfreqs which is empty
fictionfreqs <- fictionfreqs[-1,]

fictionfreqs <- as.data.frame(fictionfreqs)

#then we change the names of the columns at the top
colnames(fictionfreqs) <- c("scale", "pvalues", "difference")

#so we can see how many significant values we've obtained
print(length(which(fictionfreqs[,2] < 0.05)))

#implement the adjustment
fictionfreqs[,2] <- p.adjust(fictionfreqs[,2], method = "fdr")

#and then see how many differences are still significant
print(length(which(fictionfreqs[,2] < 0.05)))

#turn fictionfreqs into a tibble
fictionfreqs <- as_tibble(fictionfreqs)

#create new variable fictiondata to hold all rows
#in fictionfreqs where the values are lower than 0.05
fictiondata <- fictionfreqs %>%
  filter(pvalues < 0.05)

#re-create fictionfreqs so we
#have our aggregated variables
```

```
fictionfreqs <- fictionfreqs %>%
                  group_by(scale) %>%
                  dplyr::summarise(n())

#create a new variable called fictiondatums
#which holds aggregated variables
fictiondatums <- fictiondata %>%
                 group_by(scale) %>%
                 dplyr::summarise(n())

fictionfreqs <- fictionfreqs %>% filter(scale < 86)

#create numerical vector percentage which holds our
#aggregated variables expressed as percentages of each other
fictionpercentage <- (fictiondatums$`n()` /
                      fictionfreqs$`n()`) * 100

fictionnatums <- fictiondata %>%
                 group_by(scale) %>%
                 dplyr::summarise(difference = sum(difference))

fictionnatums$difference <- fictionnatums$difference /
                            fictionfreqs$`n()`

#bind them together
fictionentity <- cbind(fictionnatums,
                 fictionpercentage, "Fiction")
```

```
#replace the colnames
colnames(fictionentity) <- c("scale", "magnitude",
                              "percentage", "id")


fictionentity <- as_tibble(fictionentity)


entity <- full_join(fictionentity,
          full_join(poetryentity, dramaentity))
```

Appendix E: Modelling distances

```r
#readRDS file
data <- readRDS("fiction_frequencies.rds")

#create yearvector which is going to hold
¢all our dates so we can call on them easily
yearvector <- data$date

#create a variable which will allow us to
#switch dates around in our dataframe below
yearvector <- cbind(1:length(yearvector), yearvector)

#drop date from data
data$date <- NULL

#then we remove the first four characters from the
#colnames of data so that the colnames can be accurate
for(i in 1:length(colnames(data))) {
  colnames(data)[i] <- substr(colnames(data)[i], 5,
                        nchar(colnames(data)[i]))
}

#create data as a dataframe
data <- as.data.frame(data)

#turn it into a matrx
data <- as.matrix(data)

#apply cosine distance
data <- dist.cosine(data)

#convert our distances into a dataframe
df <- melt(as.matrix(data), varnames = c("to", "from"))

#and map our values
df$to <- mapvalues(df$to, yearvector[,1], yearvector[,2])
df$from <- mapvalues(df$from, yearvector[,1], yearvector[,2])

#convert df into a tibble
df <- as_tibble(df)

#remove all our empty distances
df <- df %>% filter(value != 0)

#divide our dataframe into distances forward
dfforward <- df %>% filter(to > from)
```

```r
#and distances backwards
dfbackward <- df %>% filter(to < from)

dfforward <- cbind(dfforward, dfforward$to - dfforward$from)

dfbackward <- cbind(dfbackward, dfbackward$from - dfbackward$to)

colnames(dfforward) <- c("to", "from", "transience", "scale")

colnames(dfbackward) <- c("to", "from", "novelty", "scale")

dfforward <- dfforward %>% filter(scale == 82)

dfbackward <- dfbackward %>% filter(scale == 82)

dfforward$to <- NULL
dfbackward$to <- NULL
dfforward$scale <- NULL
dfbackward$scale <- NULL

dates <- intersect(dfforward$from, dfbackward$from)

dfbackward <- dfbackward %>% filter(from %in% dates)
dfforward <- dfforward %>% filter(from %in% dates)

dfall <- cbind(dfbackward$from, dfbackward$novelty,
         dfforward$transience,
         dfbackward$novelty - dfforward$transience)

colnames(dfall) <- c("date", "novelty", "transience", "resonance")

dfall <- as.data.frame(dfall)

dfall$novelty <- scale(dfall$novelty)
dfall$transience <- scale(dfall$transience)
dfall$resonance <- scale(dfall$resonance)

#run our regression analysis
fit <- lm(dfall$resonance ~ dfall$novelty, data = dfall)

#cbind the residuals on the end of the dataframe
dfall <- cbind(dfall, residuals(fit))
```

Appendix F: Regularised regression

```r
#readRDS file
data <- readRDS("fiction_frequencies.rds")

#then we remove the first four characters from the colnames of
#data so that the colnames can be accurate
for(i in 1:length(colnames(data))) {
  colnames(data)[i] <- substr(colnames(data)[i], 5,
                              nchar(colnames(data)[i]))
}

#replace the first colname with Date
colnames(data)[1] <- "Date"

#create yearvector which is going to hold all our dates so
#we can call on them easily
yearvector <- data$Date
```

```
#our most pronounced break is 1837 so we need the we need the
#82 years which exist on either side of 1837, partitioned into
#two separate tibbles, pre and post
pre <- data %>% filter(Date < 1837)
pre <- pre %>% filter(Date > 1755)
post <- data %>% filter(Date > 1837)
post <- post %>% filter(Date < 1919)

#replace the date column with either zero or one
pre$Date <- 0
post$Date <- 1

pre <- as.data.frame(pre)
post <- as.data.frame(post)

#create analysis, which incorporates both pre and post
analysis <- rbind(pre, post)

#turn our date column into a factor
analysis$Date <- as.factor(analysis$Date)

#take the ID variable out of the dataframe
ID <- analysis$Date

confusionone <- 0
upwordsone <- 0
downwordsone <- 0
```

```
for(i in 1:100) {
  #create our trainining data, which randomly samples 20%
  #of our pre and post datasets
  train <- rbind(analysis[sample(which(analysis$Date == 0),
                round(0.2*length(which(analysis$Date == 0)))), ],
  #create our test data, which will incorporate everything
  #from analysis not contained in our training data
  test <- anti_join(analysis, train)
  #take our training and test data
  #from the train and test dataframes
  trainID <- train$Date
  testID <- test$Date
  #drop these columns from the training, test
  train$Date <- NULL
  test$Date <- NULL
  #we convert the data into a matrix as
  #this is the data structure glmnet needs
  train <- as.matrix(train)
  test <- as.matrix(test)
  #we then perform a cross-validated fit
  cvfit <- cv.glmnet(train, trainID, family = "binomial",
                     type.measure = "class", alpha = 0)
  #create a confusion matrix to see how often
  #one was predicted as the other and vice versa
  confusionmatrix <- confusion.glmnet(cvfit, newx = test,
                     newy = testID, s = "lambda.min")
  #extract confusionmatrix values
  confusionone <- c(confusionone, round((confusionmatrix[1]
                  + confusionmatrix[4]) /
                  sum(confusionmatrix) * 100, 2))
  analysisduplicate <- analysis
  analysisduplicate$Date <- NULL
  #re-create analysis as a matrix
  analysisduplicate <- as.matrix(analysisduplicate)
  #extract the prediction statistics
  predictions <- predict(cvfit, analysisduplicate,
                s = "lambda.min", type = "response")
  #and print the words which are positively correlated
  upword <- as_tibble(cor(predictions,
            analysisduplicate))[which(as_tibble(cor(predictions,
            analysisduplicate)) >= 0.7)]
  upwordsone <- c(upwordsone, upword)
  #and the negatively correlated words
  downword <- as_tibble(cor(predictions,
            analysisduplicate))[which(as_tibble(cor(predictions,
            analysisduplicate)) <= -0.7)]
  downwordsone <- c(downwordsone, downword)
}
```

Works Cited

Abrams, M. H. *The Mirror and the Lamp: Romantic Theory and the Critical Tradition.* Oxford University Press, 1976.

Algee-Hewitt, Mark, et al. *On Paragraphs: Scale, Themes, and Narrative Form.* Stanford Lit Lab, 2015, litlab.stanford.edu/LiteraryLabPamphlet10.pdf.

Allen, Kieran. *The Corporate Takeover of Ireland.* Irish Academic Press, 2007.

Allen, Nicholas. "Modernism and the Big House." *A History of the Modernist Novel*, edited by Gregory Castle, 2015, pp. 449–63.

Allison, Sarah, et al. "Quantitative Formalism: An Experiment." *N+1*, no. 13, Jan. 2012, nplusonemag.com/issue-13/essays/quantitative-formalism-an-experiment/.

Altieri, Charles. "*Afterword* How the 'New Modernist Studies' fails the Old Modernism." *Textual Practice*, vol. 26, no. 4, 2012, pp. 763–82, doi:10.1080/0950236X.2012.696494.

Anderson, C. W., and G. E. McMaster. "Quantification of rewriting by the Brothers Grimm: A comparison of successive versions of three tales." *Computers and the Humanities*, vol. 23, no. 4-5, 1989, pp. 341–46, doi:10.1007/BF02176639.

---. "The Emotional Tone of Foreground Lines of Poetry in Relation to Background Lines." *Literary and Linguistic Computing*, vol. 5, no. 3, 1990, pp. 226–28, doi:10.1093/llc/5.3.226.

Anderson, Perry. *Considerations on Western Marxism.* Verso, 1989.

---. *In the Tracks of Historical Materialism: The Wellek Library Lectures.* Verso, 1983.

---. "Modernity and Revolution." *New Left Review*, vol. 1, no. 144, 1984, pp. 96–113, newleftreview.org/I/144/perry-anderson-modernity-and-revolution.

---. *The Origins of Postmodernity.* Verso, 2006.

Antonia, Alexis, et al. “Language chunking, data sparseness, and the value of a long marker list: explorations with word n-grams and authorial attribution.” *Literary and Linguistic Computing*, vol. 29, no. 2, Oxford University Press, 2014, pp. 147–63, doi:10.1093/llc/fqt028.

Aoyama, H., and J. Constable. “Word length frequency and distribution in English: Part I. Prose.” *Literary and Linguistic Computing*, vol. 14, no. 3, 1999, pp. 339–58, doi:10.1093/llc/14.3.339.

Apter, Emily. *Against World Literature: On the Politics of Untranslatability.* Verso, 2013.

Argamon, Shlomo. “Interpreting Burrows’s Delta: Geometric and Probabilistic Foundations.” *Literary and Linguistic Computing*, vol. 23, no. 2, 2008, pp. 131–47, doi:10.1093/llc/fqn003.

Arnold, Taylor, and Lauren Tilton. “Distant viewing: analyzing large visual corpora.” *Digital Scholarship in the Humanities*, vol. 34, no. Supplement 1, 2019, pp. i3–16, doi:10.1093/digitalsh/fqz013.

Azim, Firdous. “Post-colonial theory.” *The Cambridge History of Literary Criticism Volume 9: Twentieth-Century Historical, Philosophical and Psychological Perspectives*, edited by Christa Knellwolf et al., 2001, pp. 235–48.

Baayen, H., et al. “Outside the cave of shadows: using syntactic annotation to enhance authorship attribution.” *Literary and Linguistic Computing*, vol. 11, no. 3, 1996, pp. 121–32, doi:10.1093/llc/11.3.121.

Baker, John Charles. “Pace: A Test of Authorship Based on the Rate at which New Words Enter an Author’s Text.” *Literary and Linguistic Computing*, vol. 3, no. 1, 1988, pp. 36–39, doi:10.1093/llc/3.1.36.

Barron, Alexander T. J., et al. “Individuals, institutions, and innovation in the debates of the French Revolution.” *Proceedings of the National Academy of Sciences of the*

United States of America, vol. 115, no. 18, National Academy of Sciences, 2018, pp. 4607–12, doi:10.1073/pnas.1717729115.

---, et al. "Supplementary Information: Individuals, Institutions, and Innovation in the Debates of the French Revolution." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 18, National Academy of Sciences, 2018, www.pnas.org/content/pnas/suppl/2018/04/16/1717729115.DCSupplemental/pnas.1717729115.sapp.pdf.

Barthélemy, Marc. "Spatial networks." *Physics Reports*, vol. 499, no. 1-3, Elsevier B.V., 2011, pp. 1–01, doi:10.1016/j.physrep.2010.11.002.

Baudelaire, Charles. *The Painter of Modern Life.* Penguin, 2010.

---. *The Poems and Prose Poems of Charles Baudelaire.* Project Gutenberg, 2011, www.gutenberg.org/files/36287/36287-h/36287-h.htm.

Beatie, Bruce A. "Measurement and the study of literature." *Computers and the Humanities*, vol. 13, no. 3, 1979, pp. 185–94, doi:10.1007/BF02395096.

Bell, Michael. "The metaphysics of Modernism." *The Cambridge Companion to Modernism*, edited by Michael Levenson, Cambridge University Press, 1999, pp. 9–32.

Benjamini, Yoav, and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society Series B (Methodological)*, vol. 57, no. 1, John Wiley & Sons, Ltd, 1995, pp. 289–300, doi:10.1111/j.2517-6161.1995.tb02031.x.

Benson, James D., and Barron Brainerd. "Chesterton's Parodies of Swinburne and Yeats: A Lexical Approach." *Literary and Linguistic Computing*, vol. 3, no. 4, 1988, pp. 226–31, doi:10.1093/llc/3.4.226.

Berman, Marshall. *All That Is Solid Melts Into Air: The Experience of Modernity.* Penguin, 1988.

Biber, Douglas. "Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation." *Literary and Linguistic Computing*, vol. 5, no. 4, 1990, pp. 257–69, doi:10.1093/llc/5.4.257.

Binongo, J., and M. Smith. "The application of principal component analysis to stylometry." *Literary and Linguistic Computing*, vol. 14, no. 4, 1999, pp. 445–66, doi:10.1093/llc/14.4.445.

Blair, Sara. "Modernism and the politics of culture." *The Cambridge Companion to Modernism*, edited by Michael Levenson, Cambridge University Press, 1999, pp. 157–73.

Bode, Katherine. *Computational Literary Studies: Participant Forum Responses.* 2019, critinq.wordpress.com/2019/04/01/computational-literary-studies-participant-forum-responses-2/.

---. *Computational Literary Studies: Participant Forum Responses, Day 2.* 2019, critinq.wordpress.com/2019/04/02/computational-literary-studies-participant-forum-responses-day-2-3/.

---. *Computational Literary Studies: Participant Forum Responses, Day 3.* 2019, critinq.wordpress.com/2019/04/03/computational-literary-studies-participant-forum-responses-day-3-2/.

Bradbury, Malcolm, and James McFarlane. "Preface." *Modernism: A Guide to European Literature 1890 - 1930*, edited by Malcolm Bradbury and James McFarlane, Penguin Books, 1991, pp. 19–56.

Brainerd, B. "The chronology of Shakespeare's plays: A statistical study." *Computers and the Humanities*, vol. 14, no. 4, 1980, pp. 221–30, doi:10.1007/BF02404431.

Brainerd, Barron. "On the distinction between a novel and a romance: A Discriminant Analysis." *Computers and the Humanities*, vol. 7, no. 5, 1973, pp. 259–70, doi:10.1007/BF02395426.

---. "Pronouns and genre in Shakespeare's drama." *Computers and the Humanities*, vol. 13, no. 1, 1979, pp. 3–16, doi:10.1007/BF02744988.

---. "Statistical analysis of lexical data using chi-squared and related distributions." *Computers and the Humanities*, vol. 9, no. 4, 1975, pp. 161–78, doi:10.1007/BF02402331.

Brandes, Ulrik, and Thomas Erlebach. "Introduction." *Network Analysis: Methodological*

Foundations, edited by Ulrik Brandes and Thomas Erlebach, Springer, 2005, pp. 1–6.

Brenkman, John. “Innovation: Notes on Nihilism and the Aesthetics of the Novel.” *The Novel: Volume 2 Forms and Themes*, edited by Franco Moretti, Princeton University Press, 2006, pp. 808–38.

Brenner, Robert. *The Boom and the Bubble: The US in the World Economy.* Verso, 2003.

Brinkmeier, Michael, and Thomas Schank. “Network Statistics.” *Network Analysis: Methodological Foundations*, edited by Ulrik Brandes and Thomas Erlebach, Springer, 2005, pp. 293–317.

Brown, Wendy. *Undoing the Demos: Neoliberalism’s Stealth Revolution.* Zone Books, 2015.

Brzezinski, M. “The New Modernist Studies: What’s Left of Political Formalism?” *The Minnesota Review*, vol. 2011, no. 76, 2011, pp. 109–25, doi:10.1215/00265667-1222083.

Burger, Peter. *Theory of the Avant-Garde.* Manchester University Press, 1984.

Burrows, J. “All the Way Through: Testing for Authorship in Different Frequency Strata.” *Literary and Linguistic Computing*, vol. 22, no. 1, 2007, pp. 27–47, doi:10.1093/llc/fqi067.

---. “’Delta’: a Measure of Stylistic Difference and a Guide to Likely Authorship.” *Literary and Linguistic Computing*, vol. 17, no. 3, 2002, pp. 267–87, doi:10.1093/llc/17.3.267.

Burrows, J. F. “‘An ocean where each kind. . .’: Statistical Analysis and Some Major Determinants of Literary Style.” *Computers and the Humanities*, vol. 23, no. 4-5, 1989, pp. 309–21, doi:10.1007/BF02176636.

---. “Modal Verbs and Moral Principles: An Aspect of Jane Austen’s Style.” *Literary and Linguistic Computing*, vol. 1, no. 1, 1986, pp. 9–23, doi:10.1093/llc/1.1.9.

---. “Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information.” *Literary and Linguistic Computing*, vol. 7, no. 2, 1992, pp. 91–109, doi:10.1093/llc/7.2.91.

---. “Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style.” *Literary and Linguistic Computing*, vol. 2, no. 2, 1987, pp. 61–70, doi:10.1093/llc/2.2.61.

Burrows, John. “Rho-grams and rho-sets: Significant links in the web of words.” *Digital Scholarship in the Humanities*, vol. 33, no. 4, 2018, pp. 724–47, doi:10.1093/llc/fqy004.

---. “Textual Analysis.” *A Companion to Digital Humanities*, edited by Susan Schreibman et al., Blackwell Publishing Ltd, 2004, pp. 323–47, doi:10.1111/b.9781405103213.2004.00026.x.

Burstein, Jessica. *Cold Modernism: Literature, Fashion, Art.* Pennsylvania State University Press, 2012.

Butler, Christopher. “Joyce the modernist.” *The Cambridge Companion to James Joyce*, edited by Derek Attridge, Cambridge University Press, 2006, pp. 67–86.

Bzdok, Danilo, et al. “Machine learning: a primer.” *Nature Methods*, Nature Publishing Group, 2017, pp. 1–2, doi:10.1038/nmeth.4526.

Caldwell, Janis McLarren. *Literature and Medicine in Nineteenth-Century Britain: From Mary Shelley to George Eliot.* Cambridge University Press, 2004.

Callinicos, Alex. “Marxism and Literary Criticism.” *The Cambridge History of Literary Criticism Volume 9: Twentieth-Century Historical, Philosophical and Psychological Perspectives*, edited by Christa Knellwolf et al., 2001, pp. 87–98.

Caruana-Galizia, Paul. “Politics and the German language: Testing Orwell’s hypothesis using the Google N-Gram corpus.” *Digital Scholarship in the Humanities*, vol. 31, no. 3, 2016, pp. 441–56, doi:10.1093/llc/fqv011.

Casanova, Pascale. *The World Republic of Letters.* Harvard University Press, 2004.

Childers, Joseph W. *Industrial culture and the Victorian novel.* Edited by Deirdre David, Cambridge University Press, 2006, pp. 77–96, doi:10.1017/CCO9780511793370.

Çiftsüren, Mehmet Nur, and Suna Akkol. “Prediction of internal egg quality characteristics and variable selection using regularization methods: ridge, LASSO and elastic net." *Archives Animal Breeding*, vol. 61, no. 3, 2018, pp. 279–84, doi:10.5194/aab-61-279-2018.

Cleary, Joe. *Outrageous Fortune: Capital and Culture in Modern Ireland.* Field Day, 2007.

---. "The World Literary System: Atlas and Epitaph." *Field Day Review*, vol. 2, 2006, pp. 196–219, doi:10.2307/30078643.

Cluett, Robert. "Style, precept, personality: A test case (Thomas Sprat, 1635–1713)." *Computers and the Humanities*, vol. 5, no. 5, 1971, pp. 257–77, doi:10.1007/BF02402207.

Coetzee, J. M. "Samuel Beckett's *Lessness:* An Exercise in Decomposition." *Computers and the Humanities*, vol. 7, no. 4, 1973, pp. 195–98, doi:10.1007/bf02403929.

Collini, Stefan. *Speaking of Universities.* Verso, 2018.

Connell, Philip. "Wordsworth's "Sonnets Dedicated to Liberty" and the British Revolutionary Past." *ELH*, vol. 85, no. 3, 2018, pp. 747–74, doi:10.1353/elh.2018.0027.

Cooper, Robert, and Mikhail Mikhailov. *Corpus Linguistics for Translation and Contrastive Studies: A guide for research.* Routledge, 2016.

Craig, D. H. "Plural Pronouns in Roman Plays by Shakespeare and Jonson." *Literary and Linguistic Computing*, vol. 6, no. 3, 1991, pp. 180–86, doi:10.1093/llc/6.3.180.

Craig, Hugh. "Contrast and Change in the Idiolects of Ben Jonson." *Computers and the Humanities*, vol. 33, no. 3, 1999, pp. 221–40, doi:10.1023/A:1002032032618.

Craik, E. M., and D. H. A. Kaferly. "The Computer and Sophocles' Trachiniae." *Literary and Linguistic Computing*, vol. 2, no. 2, 1987, pp. 86–97, doi:10.1093/llc/2.2.86.

Crane, Stephen. *Maggie: A Girl of the Streets.* Project Gutenberg, 2008, www.gutenberg.org/files/447/447-h/447-h.htm.

---. *The Red Badge of Courage: An Episode of the American Civil War.* Project Gutenberg, 2008, www.gutenberg.org/files/73/73-h/73-h.htm.

Crawley, Michael J. *Statistics: An introduction using R.* Wiley, 2015.

Da, Nan Z. “The Computational Case against Computational Literary Studies.” *Critical Inquiry*, vol. 45, no. 3, 2019, pp. 601–39, doi:10.1086/702594.

Dalen-Oskam, Karina van. “The secret life of scribes. Exploring fifteen manuscripts of Jacob van Maerlant’s Scolastica (1271).” *Literary and Linguistic Computing*, vol. 27, no. 4, 2012, pp. 355–72, doi:10.1093/llc/fqs034.

Damerau, Fred J. “The use of function word frequencies as indicators of style.” *Computers and the Humanities*, vol. 9, no. 6, 1975, pp. 271–80, doi:10.1007/BF02396290.

David, Deirdre. “Introduction.” *The Cambridge Companion to the Victorian Novel*, Cambridge University Press, 2012, pp. 1–12, doi:10.1017/CCO9780511793370.002.

Day, Gary. “Literature and the institutional context in Britain.” *The Cambridge History of Literary Criticism Volume 9: Twentieth-Century Historical, Philosophical and Psychological Perspectives*, edited by Christa Knellwolf et al., 2001, pp. 165–72.

Deleuze, Gilles, and Félix Guattari. *Anti-Oedipus: Capitalism and Schizophrenia*. Bloomsbury, 2016.

Department of Jobs, Enterprise and Innovation. *Innovation 2020: Excellence Talent Impact*. Department of Jobs, Enterprise and Innovation; Department of Jobs, Enterprise; Innovation, 2015, dbei.gov.ie/en/Publications/Publication-files/Innovation-2020.pdf.

Dews, Peter. “The Idea of Hope.” *New Left Review*, vol. 2, no. 112, 2018, pp. 99–129, newleftreview.org/issues/II112/articles/peter-dews-the-idea-of-hope.

Dickstein, Morris. “The critic and society, 1900–50.” *The Cambridge History of Literary Criticism: Volume 7. Modernism and the New Criticism*, edited by A Walton Litz et al., Cambridge, 2006.

Dixon, Peter, and David Mannion. “Goldsmith’s periodical essays: a statistical analysis of eleven doubtful cases.” *Literary and Linguistic Computing*, vol. 8, no. 1, 1993, pp. 1–19, doi:10.1093/llc/8.1.1.

Doherty, Alex. *#46 Owen Hatherley on Scott Walker’s strange musical journey*. Politics Theory Other, 2019, soundcloud.com/poltheoryother/46-owen-hatherley-on-scott-walkers-strange-musical-journey.

Drucker, Johanna. Graphesis: Visual Forms of Knowledge Production. Harvard University Press, 2014.

Eagleton, Terry. “Criticism and Politics: The Work of Raymond Williams.” *New Left Review*, vol. 1, no. 95, 1976, pp. 3–23, newleftreview.org/issues/I95/articles/terry-eagleton-criticism-and-politics-the-work-of-raymond-williams.

---. *The Ideology of the Aesthetic*. Blackwell Publishing, 1990.

Eder, Maciej. “Does size matter? Authorship attribution, small samples, big problem.” *Digital Scholarship in the Humanities*, vol. 30, no. 2, 2015, pp. 167–82, doi:10.1093/llc/fqt066.

---. “Rolling stylometry.” *Digital Scholarship in the Humanities*, vol. 31, no. 3, 2016, pp. 457–69, doi:10.1093/llc/fqv010.

Eder, Maciej, et al. “Stylometry with R: A Package for Computational Text Analysis.” *The R Journal*, Sept. 2016, pp. 107–21, journal.r-project.org/archive/2016/RJ-2016-007/RJ-2016-007.pdf.

Eder, Maciej. “Taking Stylometry to the Limits- Benchmark Study on 5,281 Texts from “Patrologia Latina”.” *Digital Humanities Conference 2015*, 2015, dh2015.org/abstracts/.

---. “Visualization in stylometry: Cluster analysis using networks.” *Digital Scholarship in the Humanities*, vol. 32, no. 1, 2017, pp. 50–64, doi:10.1093/llc/fqv061.

Elliott, Jack. “Whole genre sequencing.” *Digital Scholarship in the Humanities*, vol. 32, no. 1, 2017, pp. 65–79, doi:10.1093/llc/fqv034.

Escobar, Miguel Varela. “The Essay/ontology Workflow, Challenges in Combining Formal and Interpretive Methods.” *Literary and Linguistic Computing*, vol. 31, no. 1, 2016, pp. 84–94, doi:10.1093/llc/fqu071.

Evert, Stefan, et al. “Understanding and explaining Delta measures for authorship attribution.” *Digital Scholarship in the Humanities*, vol. 32, no. 2, 2017, pp. ii4–16, doi:10.1093/llc/fqx023.

Eysteinsson, Ástrádur. *The Concept of Modernism.* Cornell University Press, 1990.

Febvre, Lucien, and Henri-Jean Martin. *The Coming of the Book: The Impact of Printing, 1450 - 1800.* Edited by Geoffrey Nowell-Smith and David Wootton, Verso, 2010.

Feng, Haoda. "Quantitative Corpus Linguistics with R: A Practical Introduction (Second edition). Stefan Th. Gries." *Digital Scholarship in the Humanities*, vol. 34, no. 3, 2019, pp. 696–98, doi:10.1093/llc/fqz021.

Fletcher, John, and Malcolm Bradbury. "The Introverted Novel." *Modernism: A Guide to European Literature 1890 - 1930*, edited by Malcolm Bradbury and James McFarlane, Penguin Books, 1991, pp. 394–415.

Flint, Kate. "The Victorian novel and its readers." *The Cambridge Companion to the Victorian Novel*, edited by Deirdre David, Cambridge University Press, 2006, pp. 17–36, doi:10.1017/CCOL0521641500.002.

Flood, Alison. "Scientists find evidence of mathematical structures in classic books." *The Guardian*, 2016, www.theguardian.co.uk/books/2016/jan/27/scientists-reveal-multifractal-structure-of-finnegans-wake-james-joyce.

Foley, John Miles. "A computer analysis of metrical patterns in Beowulf." *Computers and the Humanities*, vol. 12, no. 1-2, 1978, pp. 71–80, doi:10.1007/BF02392918.

Forsyth, R. S., et al. "Cicero, Sigonio, and Burrows: investigating the authenticity of the Consolatio." *Literary and Linguistic Computing*, vol. 14, no. 3, 1999, pp. 375–400, doi:10.1093/llc/14.3.375.

Forsyth, R. S., and D. I. Holmes. "Feature-finding for text classification." *Digital Scholarship in the Humanities*, vol. 11, no. 4, 1996, pp. 163–74, doi:10.1093/llc/11.4.163.

Forsyth, Richard S. "Stylochronometry with substrings, or: a poet young and old." *Literary and Linguistic Computing*, vol. 14, no. 4, 1999, pp. 467–78, doi:10.1093/llc/14.4.467.

Forsyth, Richard S., and Phoenix W. Y. Lam. "Found in translation: To what extent is authorial discriminability preserved by translators?" *Literary and Linguistic*

Computing, vol. 29, no. 2, Oxford University Press, 2014, pp. 199–217, doi:10.1093/llc/fqt018.

Fortier, P. A. “Theory, Methods and Applications: Some Examples in French Literature.” *Literary and Linguistic Computing*, vol. 6, no. 3, 1991, pp. 192–96, doi:10.1093/llc/6.3.192.

Fortier, Paul A. “Some statistics of themes in the French novel.” *Computers and the Humanities*, vol. 23, no. 4-5, 1989, pp. 293–99, doi:10.1007/BF02176634.

Fortunato, Santo, and Darko Hric. “Community detection in networks: A user guide.” *Physics Reports*, vol. 659, Elsevier B.V., 2016, pp. 1–44, doi:10.1016/j.physrep.2016.09.002.

Galsworthy, John. *The Burning Spear*. Project Gutenberg, 2018, www.gutenberg.org/files/2905/2905-h/2905-h.htm.

Garcia-Zorita, Carlos, and Ana R. Pacios. “Topic modelling characterization of Mudejar art based on document titles.” *Digital Scholarship in the Humanities*, vol. 33, no. 3, 2018, pp. 529–39, doi:10.1093/llc/fqx055.

Gladwin, Alexander A. G., et al. “Stylometry and collaborative authorship: Eddy, Lovecraft, and ‘The Loved Dead’.” *Digital Scholarship in the Humanities*, vol. 32, no. 1, Oxford University Press, 2017, pp. 123–40, doi:10.1093/llc/fqv026.

Gooding, P. “Mass digitization and the garbage dump: The conflicting needs of quantitative and qualitative methods.” *Literary and Linguistic Computing*, vol. 28, no. 3, 2013, pp. 425–31, doi:10.1093/llc/fqs054.

Greenspan, Brian. “Are Digital Humanists Utopian?” *Debates in the Digital Humanities 2016*, edited by Matthew K Gold and Lauren F Klein, University of Minnesota Press, 2016, pp. 393–409.

Greenwood, H. H. “St Paul Revisited–a Computational Result.” *Literary and Linguistic Computing*, vol. 7, no. 1, 1992, pp. 43–47, doi:10.1093/llc/7.1.43.

---. “St Paul Revisited–Word Clusters in Multidimensional Space.” *Literary and Linguistic Computing*, vol. 8, no. 4, 1993, pp. 211–19, doi:10.1093/llc/8.4.211.

Grieve, J. “Quantitative Authorship Attribution: An Evaluation of Techniques.” *Literary and Linguistic Computing*, vol. 22, no. 3, 2007, pp. 251–70, doi:10.1093/llc/fqm020.

Guo, Shesen, et al. “Distribution of English syllables in e-books of Project Gutenberg and the evolution of syllable number in two subcorpora.” *Digital Scholarship in the Humanities*, vol. 30, no. 3, 2015, pp. 344–53, doi:10.1093/llc/fqu013.

Gurney, Penelope J., and Lyman W. Gurney. “Authorship Attribution of the Scriptores Historiae Augustae.” *Literary and Linguistic Computing*, vol. 13, no. 3, 1998, pp. 119–31, doi:10.1093/llc/13.3.119.

---. “Subsets and Homogeneity: Authorship Attribution in the Scriptories Historiae Augustae.” *Literary and Linguistic Computing*, vol. 13, no. 3, 1998, pp. 133–40, doi:10.1093/llc/13.3.133.

Guy, Josephine M., and Ian Small. “The British ’man of letters’ and the rise of the professional.” *The Cambridge History of Literary Criticism: Volume 7. Modernism and the New Criticism*, edited by A Walton Litz et al., Cambridge, 2006.

Habermas, Jürgen. *The Philosophical Discourse of Modernity: Twelve Lectures.* Polity Press, 1987.

Habib, M. A. R. *A History of Literary Criticism and Theory: From Plato to the Present.* Blackwell Publishing, 2008.

Hawkes, Terence. *Structuralism and Semiotics.* Methuen & Co., 1978.

Hegel, George Wilhelm Friedrich. *Aesthetics: Lectures on Fine Art.* Clarendon Press, 1988.

Henkels Jr, Robert M., and Esteban R. Egea. “Using a computer-generated concordance to analyze and document stylistic devices in Robert Pinget’s fable.” *Computers and the Humanities*, vol. 11, no. 6, 1977, pp. 325–38, doi:10.1007/BF02428133.

Heuser, Ryan, and Long Le-Khac. “A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method.” *Stanford Lit Lab*, 2012, pp. 1–68, litlab.stanford.edu/LiteraryLabPamphlet4.pdf.

Hilton, Michael L., and David I. Holmes. “An Assessment of Cumulative Sum Charts for

Authorship Attribution." *Literary and Linguistic Computing*, vol. 8, no. 2, 1993, pp. 73–80, doi:10.1093/llc/8.2.73.

Hirst, Graeme, and Ol'ga Feiguina. "Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts." *Literary and Linguistic Computing*, vol. 22, no. 4, 2007, pp. 405–17, doi:10.1093/llc/fqm023.

Hobsbawm, Eric. *The Age of Revolution 1789 - 1848.* Abacus, 2010.

Hoerl, Arthur E., and Robert W. Kennard. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics*, vol. 12, no. 1, 1970, pp. 55–67, doi:10.1080/00401706.1970.10488634.

Hogenraad, Robert, et al. "The enemy within: Autocorrelation bias in content analysis of narratives." *Computers and the Humanities*, vol. 30, no. 6, 1997, pp. 433–39, doi:10.1007/BF00057939.

Hogue, Stephane, and Alastair McKinnon. "Uforanderlige and Uforanderlighed: More about their Differences." *Literary and Linguistic Computing*, vol. 2, no. 2, 1987, pp. 98–107, doi:10.1093/llc/2.2.98.

Holmes, D. I. "The Evolution of Stylometry in Humanities Scholarship." *Literary and Linguistic Computing*, vol. 13, no. 3, 1998, pp. 111–17, doi:10.1093/llc/13.3.111.

Holmes, D. I., and D. W. Crofts. "The diary of a public man: a case study in traditional and non-traditional authorship attribution." *Literary and Linguistic Computing*, vol. 25, no. 2, 2010, pp. 179–97, doi:10.1093/llc/fqq005.

Hoorn, Johann F., et al. "Neural network identification of poets using letter sequences." *Literary and Linguistic Computing*, vol. 14, no. 3, 1999, pp. 311–38, doi:10.1093/llc/14.3.311.

Hoover, David L. "Delta Prime?" *Literary and Linguistic Computing*, vol. 19, no. 4, 2004, pp. 477–95, doi:10.1093/llc/19.4.477.

---. "Frequent Collocations and Authorial Style." *Literary and Linguistic Computing*, vol. 18, no. 3, 2003, pp. 261–86, doi:10.1093/llc/18.3.261.

---. "Frequent Word Sequences and Statistical Stylistics." *Literary and Linguistic*

Computing, vol. 17, no. 2, 2002, pp. 157–80, doi:10.1093/llc/17.2.157.

---. “Multivariate Analysis and the Study of Style Variation.” *Literary and Linguistic Computing*, vol. 18, no. 4, 2003, pp. 341–60, doi:10.1093/llc/18.4.341.

---. “Simulations and difficult problems.” *Digital Scholarship in the Humanities*, vol. 34, no. 4, 2019, pp. 874–92, doi:10.1093/llc/fqz034.

---. “Statistical Stylistics and Authorship Attribution: an Empirical Investigation.” *Literary and Linguistic Computing*, vol. 16, no. 4, 2001, pp. 421–44, doi:10.1093/llc/16.4.421.

---. “Testing Burrows’s Delta.” *Literary and Linguistic Computing*, vol. 19, no. 4, 2004, pp. 453–75, doi:10.1093/llc/19.4.453.

Hoover, David L., and Shervin Hess. “An exercise in non-ideal authorship attribution: the mysterious Maria Ward.” *Literary and Linguistic Computing*, vol. 24, no. 4, 2009, pp. 467–89, doi:10.1093/llc/fqp027.

Hou, Renkui, and Minghu Jiang. “Analysis on Chinese quantitative stylistic features based on text mining.” *Digital Scholarship in the Humanities*, vol. 31, no. 2, 2016, pp. 357–67, doi:10.1093/llc/fqu067.

Ilsemann, Hartmut. “Forensic stylometry.” *Digital Scholarship in the Humanities*, 2018, pp. 1–15, doi:10.1093/llc/fqy023.

---. “Stylometry approaching Parnassus.” *Digital Scholarship in the Humanities*, vol. 33, no. 3, 2018, pp. 548–56, doi:10.1093/llc/fqx058.

Inwood, Michael. “Introduction.” *Introductory Lectures on Aesthetics*, edited by Michael Inwood, Penguin, 2004.

Irizarry, Estelle. “One Writer, Two Authors: Resolving the Polemic of Latin America’s First Published Novel.” *Literary and Linguistic Computing*, vol. 6, no. 3, 1991, pp. 175–79, doi:10.1093/llc/6.3.175.

---. “The two authors of Columbus’Diary.” *Computers and the Humanities*, vol. 27, no. 2, 1993, pp. 85–92, doi:10.1007/BF01830301.

Jameson, Frederic. *A Singular Modernity: Essay on the Ontology of the Present.* Verso, 2012.

---. *Marxism and Form: Twentieth-Century Dialectical Theories of Literature.* Princeton University Press, 1974.

---. *The Hegel Variations: On the Phenomenology of Spirit.* Verso, 2017.

---. *The Modernist Papers.* Verso, 2007.

---. *The Political Unconscious: Narrative as a Socially Symbolic Act.* Routledge, 2002.

---. *The Prison House of Language: A Critical Account of Structuralism and Russian Formalism.* Princeton University Press, 1974.

Jameson, Fredric. *The Ancients and the Postmoderns: On the Historicity of Forms.* Verso, 2017.

---. *Valences of the Dialectic.* Verso, 2009.

Jannidis, Fotis, et al. "Improving Burrows' Delta: An empirical evaluation of text distance measures." *Digital Humanities Conference 2015*, 2015, www.researchgate.net/publication/280086768_Improving_An_empirical_evaluation_of_text_distance_measures/link/573ad8ae08ae9f741b2d3d40/download.

Jockers, Matthew L., and Daniela M. Witten. "A comparative study of machine learning methods for authorship attribution." *Literary and Linguistic Computing*, vol. 25, no. 2, 2010, pp. 215–23, doi:10.1093/llc/fqq001.

Joliffe, I. T. *Principal Components Analysis.* Springer, 2004.

Joyce, Simon. *Modernism and Naturalism in British and Irish Fiction, 1880–1930.* Cambridge University Press, 2015, doi:10.1017/CBO9781316018668.

---. *The Victorians in the Rearview Mirror.* Ohio University Press, 2007.

Juola, Patrick. "The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions." *Digital Scholarship in the Humanities*, vol. 30, no. Supplement 1, 2015, pp. i100–13, doi:10.1093/llc/fqv040.

Kalaidjian, Walter. *The Cambridge Companion to American Modernism.* Cambridge University Press, 2005.

Kennedy, Sinéad. *Marxism After Modernism: Anglo-American Leftist Theorisations of Modernism in the Later Twentieth Century.*

2007, mural.maynoothuniversity.ie/5282/1/Sinead_Kennedy_20140722154957.pdf.

Khmelev, Dmitri V., and Fiona J. Tweedie. “Using Markov chains for identification of writers.” *Literary and Linguistic Computing*, vol. 16, no. 3, 2001, pp. 299–307, doi:10.1093/llc/16.3.299.

Kjell, Bradley. “Authorship Determination Using Letter Pair Frequency Features with Neural Network Classifiers.” *Literary and Linguistic Computing*, vol. 9, no. 2, 1995, pp. 119–24, doi:10.1093/llc/9.2.119.

Klein, Lauren F., and Matthew K. Gold. *Debates in the Digital Humanities 2016*. Edited by Matthew K Gold and Lauren F Klein, University of Minnesota Press, 2016.

Knox, T. M. “Translator’s Preface.” *Aesthetics: Lectures on Fine Art*, Clarendon Press, 1988.

Kocher, Mirco, and Jacques Savoy. “Distributed language representation for authorship attribution.” *Digital Scholarship in the Humanities*, vol. 33, no. 2, 2018, pp. 425–41, doi:10.1093/llc/fqx046.

Kozima, Hideki, and Teiji Furugori. “Segmenting narrative text into coherent scenes.” *Literary and Linguistic Computing*, vol. 9, no. 1, 1994, pp. 13–19, doi:10.1093/llc/9.1.13.

Kullback, S., and R. A. Leibler. “On Information and Statistics.” *The Annals of Mathematical Statistics*, vol. 22, no. 1, 1951, pp. 79–86.

Laan, Nancy M. “Stylometry and Method: The Case of Euripides.” *Literary and Linguistic Computing*, vol. 10, no. 4, 1995, pp. 271–78, doi:10.1093/llc/10.4.271.

Laffal, Julius. “A concept analysis of Jonathan Swift’s ”A Tale of a Tub” and ”Gulliver’s Travels”.” *Computers and the Humanities*, vol. 29, no. 5, 1995, pp. 339–61, doi:10.1007/BF02279526.

Landow, George P. “Hypertext in literary education, criticism, and scholarship.” *Computers and the Humanities*, vol. 23, no. 3, 1989, pp. 173–98, doi:10.1007/BF00056142.

Lantz, Brett. *Machine Learning with R*. Packt Publishing, 2013.

Lawrence, D. H. *The Captain's Doll.* Project Gutenberg Australia, 2002, gutenberg.net.au/ebooks02/0200841h.html.

Lee, Changsoo. "Do language combinations affect translators' stylistic visibility in translated texts?" *Digital Scholarship in the Humanities*, vol. 33, no. 3, 2018, pp. 592–603, doi:10.1093/llc/fqx056.

Lee, Hermione. *Virginia Woolf.* Vintage, 1997.

Lehan, Richard. "The European Background." *The Cambridge Companion to American Realism and Naturalism: From Howells to London*, edited by Donald Pizer, Cambridge University Press, 1995, pp. 47–74.

Lenin, Vladimir Ilyich. *Karl Marx: A Brief Biographical Sketch with an Exposition of Marxism.* Foreign Languages Press, 1970.

Lessard, Greg, and Johanne Bénard. "Computerizing Céline." *Computers and the Humanities*, vol. 27, no. 5-6, 1993, pp. 387–94, doi:10.1007/BF01829389.

Lessard, Gregory, and Jean-Jacques Hamm. "Computer-aided Analysis of Repeated Structures: the Case of Stendhal's *Armance*." *Literary and Linguistic Computing*, vol. 6, no. 4, 1991, pp. 246–52, doi:10.1093/llc/6.4.246.

Levenson, J. C. "*The Red Badge of Courage* and *McTeague*: Passage to Modernity." *The Cambridge Companion to American Realism and Naturalism: From Howells to London*, edited by Donald Pizer, Cambridge University Press, 1995, pp. 154–77.

Lever, Jake, et al. "Points of Significance: Regularization." *Nature Methods*, vol. 13, no. 10, Nature Publishing Group, 2016, pp. 1–2, doi:10.1038/nmeth.4014.

Liedman, Sven-Eric. *A World to Win: The Life and Works of Karl Marx.* Verso, 2018.

Lijffijt, Jefrey, et al. "Significance testing of word frequencies in corpora." *Digital Scholarship in the Humanities*, vol. 31, no. 2, 2016, pp. 374–97, doi:10.1093/llc/fqu064.

Logan, Harry M., and Grace B. Logan. "The Case of the Canterbury Pilgrims: Sentence Semantics and World View in Frag. 1 of The Canterbury Tales." *Literary and Linguistic Computing*, vol. 5, no. 3, 1990, pp. 242–47, doi:10.1093/llc/5.3.242.

MacKay, Marina. *Modernism and World War II.* Cambridge University Press, 2007.

Mannion, David, and Peter Dixon. "Sentence-length and Authorship Attribution: the Case of Oliver Goldsmith." *Literary and Linguistic Computing*, vol. 19, no. 4, 2004, pp. 497–508, doi:10.1093/llc/19.4.497.

Mansell, Darrell. "*The Old Man and the Sea* and the computer." *Computers and the Humanities*, vol. 8, no. 4, 1974, pp. 195–206, doi:10.1007/BF02402341.

Mao, Douglas, and Rebecca L. Walkowitz. "Introduction." *Bad Modernisms*, edited by Rebecca L Walkowitz and Douglas Mao, Duke University Press, 2006, pp. 1–18.

---. "The New Modernist Studies." *PMLA*, vol. 123, no. 3, 2008, pp. 737–48, doi:10.1632/pmla.2008.123.3.737.

Marcus, Mitchell P., et al. "Building a Large Annotated Corpus of English - The Penn Treebank." *Comput. Linguistics*, 1993, repository.upenn.edu/cis_reports/237/.

Martindale, Colin, and Dean McKenzie. "On the utility of content analysis in author attribution: *The Federalist*." *Computers and the Humanities*, vol. 29, no. 4, 1995, pp. 259–70, doi:10.1007/BF01830395.

Marx, Karl. *Capital: A Critical Analysis of Capitalist Production.* Wordsworth Editions, 2013.

Masuda, Naoki, et al. "Random walks and diffusion on networks." *Physics Reports*, vol. 716-717, Elsevier B.V., 2017, pp. 1–58, doi:10.1016/j.physrep.2017.07.007.

Matthews, Robert A. J., and Thomas V. N. Merriam. "Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher." *Literary and Linguistic Computing*, vol. 8, no. 4, 1993, pp. 203–09, doi:10.1093/llc/8.4.203.

Mazzucato, Mariana. *Mission-Oriented Research & Innovation in the European Union: A problem-solving approach to fuel innovation-led growth.* European Commission, 2018, ec.europa.eu/info/sites/info/files/mazzucato_report_2018.pdf.

McArthur, Tom, et al. *The Oxford Companion to the English Language.* Edited by Tom McArthur et al., vol. 1, Oxford University Press, 2018, doi:10.1093/acref/9780199661282.001.0001.

McColly, William B. “Style and structure in the Middle English poem *Cleanness*.” *Computers and the Humanities*, vol. 21, no. 3, 1987, pp. 169–76, doi:10.1007/BF02252793.

McDonald, Gary C. “Ridge regression.” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 1, John Wiley & Sons, Ltd, 2009, pp. 93–100, doi:10.1002/wics.14.

McFarlane, James. “The Mind of Modernism.” *Modernism: A Guide to European Literature 1890 - 1930*, edited by Malcolm Bradbury and James McFarlane, Penguin Books, 1991, pp. 71–94.

McKenna, W., and A. Antonia. “’A Few Simple Words’ of Interior Monologue in Ulysses: Reconfiguring the Evidence.” *Literary and Linguistic Computing*, vol. 11, no. 2, 1996, pp. 55–66, doi:10.1093/llc/11.2.55.

Menand, Louis, and Lawrence Rainey. “Introduction.” *The Cambridge History of Literary Criticism: Volume 7. Modernism and the New Criticism*, edited by A Walton Litz et al., Cambridge, 2006.

Mengham, Rod. “Postwar modernism in the 1920s and 1930s: The mammoth in the basement.” *The Cambridge Companion to the Twentieth-Century English Novel*, edited by Robert L. Caseiro, Cambridge University Press, 2009, pp. 71–88, doi:10.1017/CCOL9780521884167.006.

Merriam, Thomas. “An Experiment with the Federalist Papers.” *Computers and the Humanities*, vol. 23, no. 3, 1989, pp. 251–54, doi:10.1007/BF00056147.

Merriam, Thomas V. N., and Robert A. J. Matthews. “Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe.” *Literary and Linguistic Computing*, vol. 9, no. 1, 1994, pp. 1–6, doi:10.1093/llc/9.1.1.

Mesthene, Emmanuel G. “Technology and humanistic values.” *Computers and the Humanities*, vol. 4, no. 1, 1969, pp. 1–10, doi:10.1007/BF02393443.

Milic, Louis. “Progress in stylistics: Theory, statistics, computers.” *Computers and the Humanities*, vol. 25, no. 6, 1991, pp. 393–400, doi:10.1007/BF00141189.

Milic, Louis T. “The Next Step.” *Computers and the Humanities*, vol. 1, no. 1, 1966, pp. 3–6, doi:10.1007/BF00188010.

Milic, Louis Tonko. *A Quantitative Approach to the Style of Jonathan Swift.* Mouton & Co., 1967.

Moi, Toril. *Henrik Ibsen and the Birth of Modernism: Art, Theatre, Philosophy.* Oxford University Press, 2006.

Moisl, Hermann. *Cluster Analysis for Corpus Linguistics.* Edited by Reinhard Köhler et al., De Gruyter Mouton, 2015.

Moretti, Franco. *Atlas of the European Novel 1800 - 1900.* Verso, 1998.

---. *Distant Reading.* Verso, 2013.

---. *Graphs, Maps, Trees: Abstract Models for a Literary History.* Verso, 2007.

---. *The Bourgeois: Between History and Literature.* Verso, 2013.

---. “The Roads to Rome: *Literary Studies, Hermeneutics, Quantification.*” *New Left Review*, vol. 2, no. 124, 2020, pp. 125–36, newleftreview.org/issues/II124/articles/franco-moretti-the-roads-to-rome.pdf.

Mulhern, Francis. “Inconceivable History: Hyperphasia and Disavowal.” *The Novel: Volume 2 Forms and Themes*, edited by Franco Moretti, Princeton University Press, 2006, pp. 777–807.

Nadel, Ira B., and Stephen N. Matsuba. “The ’Cunning Pattern of Excelling Nature’: Shakespeare and the Literary Application of DISCAN.” *Literary and Linguistic Computing*, vol. 5, no. 3, 1990, pp. 229–34, doi:10.1093/llc/5.3.229.

Nolan, Emer. *Catholic Emancipations: Irish Fiction from Thomas Moore to James Joyce.* Syracuse University Press, 2007.

North, Joseph. *Literary Criticism: A Political History.* Harvard University Press, 2017.

Oakes, Michael P. “Computer stylometry of C. S. Lewis’s The Dark Tower and related texts.” *Digital Scholarship in the Humanities*, vol. 33, no. 3, 2018, pp. 637–50, doi:10.1093/llc/fqx043.

Ogutu, Joseph O., et al. “Genomic selection using regularized linear regression models:

Ridge regression, lasso, elastic net and their extensions." *BMC Proceedings*, vol. 6, no. SUPPL 2, BioMed Central, 2012, pp. 1–6, doi:10.1186/1753-6561-6-S2-S10.

Olsen, Mark. "Critical theory and textual computing: Comments and suggestions." *Computers and the Humanities*, vol. 27, no. 5-6, 1993, pp. 395–400, doi:10.1007/BF01829390.

Oostdijk, N. "The Language of Dialogue in Fiction." *Literary and Linguistic Computing*, vol. 5, no. 3, 1990, pp. 235–41, doi:10.1093/llc/5.3.235.

Opas, Lisa Lena, and Fiona Tweedie. "The magic carpet ride: reader involvement in romantic fiction." *Literary and Linguistic Computing*, vol. 14, no. 1, 1999, pp. 89–101, doi:10.1093/llc/14.1.89.

Oser, Lee. *The Ethics of Modernism: Moral Ideas in Yeats, Eliot, Joyce, Woolf, and Beckett.* Cambridge University Press, 2007.

Pasanek, B., and D. Sculley. "Mining millions of metaphors." *Literary and Linguistic Computing*, vol. 23, no. 3, 2008, pp. 345–60, doi:10.1093/llc/fqn010.

Pearl, Lisa, et al. "The character in the letter: Epistolary attribution in Samuel Richardson's Clarissa." *Digital Scholarship in the Humanities*, vol. 32, no. 2, 2017, pp. 355–76, doi:10.1093/llc/fqw007.

Peer, W. van. "Quantitative studies of literature. A critique and an outlook." *Computers and the Humanities*, vol. 23, no. 4-5, 1989, pp. 301–07, doi:10.1007/BF02176635.

Pinkard, Terry. *Hegel: A Biography.* Cambridge University Press, 2000.

Pinkney, Tony. "Understanding Modernism: A Response to Franco Moretti." *New Left Review*, vol. 1, no. 167, 1988, pp. 124–27, newleftreview.org/issues/I167/articles/tony-pinkney-understanding-modernism-a-response-to-franco-moretti.pdf.

Piper, Andrew. *Enumerations: Data and Literary Study.* University of Chicago Press, 2018.

Pizer, Donald. "Introduction." *The Cambridge Companion to American Realism and Naturalism: From Howells to London*, edited by Donald Pizer, Cambridge University

Press, 1995, pp. 1–18.

Plant, Raymond. “Hegel and Political Economy—I.” *New Left Review*, vol. 1, no. 103, 1977, pp. 79–92, newleftreview.org/issues/I103/articles/raymond-plant-hegel-and-political-economy-part-i.

---. “Hegel and Political Economy—II.” *New Left Review*, vol. 1, no. 104, 1977, pp. 103–13, newleftreview.org/issues/I104/articles/raymond-plant-hegel-and-political-economy-part-ii.

Potter, Rosanne G. “Literary criticism and literary computing: The difficulties of a synthesis.” *Computers and the Humanities*, vol. 22, no. 2, 1988, pp. 91–97, doi:10.1007/BF00057648.

---. “Toward a syntactic differentiation of period style in modern drama: Significant between-play variability in 21 english-language plays.” *Computers and the Humanities*, vol. 14, no. 3, 1980, pp. 187–96, doi:10.1007/BF02403767.

Prendergast, Christopher. “Introduction.” *Debating World Literature*, edited by Christopher Prendergast, Verso, 2004, pp. vii–xiii.

Radiolab. *Vanishing Words.* WNYC Studios, 2010, www.wnycstudios.org/podcasts/radiolab/articles/91960-vanishing-words.

Rainey, Lawrence. “The cultural economy of Modernism.” *The Cambridge Companion to Modernism*, edited by Michael Levenson, Cambridge University Press, 1999, pp. 33–69.

Ramsay, Stephen. *Reading Machines: Towards an Algorithmic Criticism.* University of Illinois Press, 2011.

Readings, Bill. *The University in Ruins.* Harvard University Press, 1998.

Richardson, Leonard, and James English. “Project Gutenberg Books Are Real.” *The Journal of Electronic Publishing*, vol. 18, no. 1, 2015, doi:10.3998/3336451.0018.126.

Rizvi, Pervez. “An improvement to Zeta.” *Digital Scholarship in the Humanities*, vol. 34, no. 2, 2019, pp. 419–22, doi:10.1093/llc/fqy039.

Roberts, Alan. “Rhythm in Prose and the Serial Correlation of Sentence Lengths: a

Joyce Cary Case Study." *Literary and Linguistic Computing*, vol. 11, no. 1, 1996, pp. 33–39, doi:10.1093/llc/11.1.33.

Robey, David. "Sound and Sense in the Divine Comedy." *Literary and Linguistic Computing*, vol. 2, no. 2, 1987, pp. 108–15, doi:10.1093/llc/2.2.108a.

Rockwell, Geoffrey. "What is Text Analysis, Really?" *Literary and Linguistic Computing*, vol. 18, no. 2, 2003, pp. 209–19, doi:10.1093/llc/18.2.209.

Rose, Gillian. *The Melancholy Science: An Introduction to the Thought of Theodor W. Adorno*. Verso, 2014.

Rybicki, Jan, and Maciej Eder. "Deeper Delta across genres and languages: do we really need the most frequent words?" *Literary and Linguistic Computing*, vol. 26, no. 3, Oxford University Press, 2011, pp. 315–21, doi:10.1093/llc/fqr031.

Rybicki, Jan, and Magda Heydel. "The stylistics and stylometry of collaborative translation: Woolf's Night and Day in Polish." *Literary and Linguistic Computing*, vol. 28, no. 4, 2013, pp. 708–17, doi:10.1093/llc/fqt027.

Saccenti, Edoardo, and Leonardo Tenori. "Multivariate modeling of the collaboration between Luigi Illica and Giuseppe Giacosa for the librettos of three operas by Giacomo Puccini." *Digital Scholarship in the Humanities*, vol. 30, no. 3, 2014, pp. 405–22, doi:10.1093/llc/fqu006.

Said, Edward. "Introduction." *Mimesis: The Representation of Reality in Western Literature*, Princeton University Press, 2013.

Sainte-Marie, Paule, et al. "An application of principal component analysis to the works of Molière." *Computers and the Humanities*, vol. 7, no. 3, 1973, pp. 131–37, doi:10.1007/BF02403851.

Savoy, Jacques. "Comparative evaluation of term selection functions for authorship attribution." *Digital Scholarship in the Humanities*, vol. 30, no. 2, 2015, pp. 246–61, doi:10.1093/llc/fqt047.

Sayoud, Halim. "Author discrimination between the Holy Quran and Prophet's statements." *Literary and Linguistic Computing*, vol. 27, no. 4, 2012, pp. 427–44, doi:10.1093/llc/fqs014.

Schaalje, G. Bruce, et al. "Extended nearest shrunken centroid classification: A new method for open-set authorship attribution of texts of varying sizes." *Literary and Linguistic Computing*, vol. 26, no. 1, 2011, pp. 71–88, doi:10.1093/llc/fqq029.

Schöberlein, Stefan. "Poe or Not Poe? A Stylometric Analysis of Edgar Allan Poe's Disputed Writings." *Digital Scholarship in the Humanities*, vol. 32, no. 3, 2017, pp. 643–59, doi:10.1093/llc/fqw019.

Schreibman, Susan, et al. "The Digital Humanities and Humanities Computing: An Introduction." *A Companion to Digital Humanities*, edited by Susan Schreibman et al., Blackwell Publishing Ltd, 2004, pp. 3–19, doi:10.1111/b.9781405103213.2004.00004.x.

Scofield, Martin. *The Cambridge Introduction to The American Short Story*. Cambridge University Press, 2006.

Sinclair, Stéfan. "Computer-Assisted Reading: Reconceiving Text Analysis." *Literary and Linguistic Computing*, vol. 18, no. 2, 2003, pp. 175–84, doi:10.1093/llc/18.2.175.

Smith, J. A., and C. Kelly. "Stylistic Constancy and Change across Literary Corpora: Using Measures of Lexical Richness to Date Works." *Computers and the Humanities*, vol. 36, no. 4, 2002, pp. 411–30.

Smith, M. W. A. "A Critical Review of Word-links as a Method for Investigating Shakespearean Chronology and Authorship." *Literary and Linguistic Computing*, vol. 1, no. 4, 1986, pp. 202–06, doi:10.1093/llc/1.4.202.

---. "An investigation of Morton's method to distinguish Elizabethan playwrights." *Computers and the Humanities*, vol. 19, no. 1, 1985, pp. 3–21, doi:10.1007/BF02259614.

---. "Hapax Legomena in Prescribed Positions: An Investigation of Recent Proposals to Resolve Problems of Authorship." *Literary and Linguistic Computing*, vol. 2, no. 3, 1987, pp. 145–52, doi:10.1093/llc/2.3.145.

---. "The authorship of Acts I and II of Pericles: a new approach using first words of speeches." *Computers and the Humanities*, vol. 22, no. 1, 1988, pp. 23–41, doi:10.1007/BF00056347.

Smith, Peter W. H., and W. Aldridge. “Improving Authorship Attribution: Optimizing Burrows’ Delta Method*.” *Journal of Quantitative Linguistics*, vol. 18, no. 1, 2011, pp. 63–88, doi:10.1080/09296174.2011.533591.

Stamou, Constantina. “Stylochronometry: Stylistic Development, Sequence of Composition, and Relative Dating.” *Literary and Linguistic Computing*, vol. 23, no. 2, 2008, pp. 181–99, doi:10.1093/llc/fqm029.

Stanikunas, Daumantas, et al. “Comparison of distance and similarity measures for stylometric analysis of Lithuanian texts.” *International Conference for Young Researchers in Informatics, Mathematics and Engineering*, 2015, ceur-ws.org/Vol-1852/p01.pdf.

Stapel, Rombert J. “Layer on layer. ’Computational archaeology’ in 15th-century Middle Dutch historiography.” *Literary and Linguistic Computing*, vol. 28, no. 2, 2013, pp. 344–58, doi:10.1093/llc/fqs046.

Stratil, M., and R. J. Oakley. “A Disputed Authorship Study of Two Plays Attributed to Tirso de Molina.” *Literary and Linguistic Computing*, vol. 2, no. 3, 1987, pp. 153–60, doi:10.1093/llc/2.3.153.

Sula, Chris Alen, and Heather V. Hill. “The early history of digital humanities: An analysis of Computers and the Humanities (1966–2004) and Literary and Linguistic Computing (1986–2004).” *Digital Scholarship in the Humanities*, vol. 34, no. Supplement 1, 2019, pp. i190–206, doi:10.1093/llc/fqz072.

Sutherland, Kathryn. “A Guide Through the Labyrinth: Dickens’s *Little Dorrit* as Hypertext.” *Literary and Linguistic Computing*, vol. 5, no. 4, 1990, pp. 305–09, doi:10.1093/llc/5.4.305.

Svensson, Patrik. “Three Premises of Big Digital Humanities.” *Big Digital Humanities*, University of Michigan Press, 2016, pp. 82–130, doi:10.2307/j.ctv65sx0t.7.

Temple, J. T. “A Multivariate Synthesis of Published Platonic Stylometric Data.” *Literary and Linguistic Computing*, vol. 11, no. 2, 1996, pp. 67–75, doi:10.1093/llc/11.2.67.

The Stanford Lit Lab. "Style at the Scale of the Sentence." *N+1*, no. 17, 2013, nplusonemag.com/issue-17/essays/style-at-the-scale-of-the-sentence/.

Thoiron, Philippe. "Diversity index and entropy as measures of lexical richness." *Computers and the Humanities*, vol. 20, no. 3, 1986, pp. 197–202, doi:10.1007/BF02404461.

Tse, Emily K., et al. "Unravelling the Purple Thread: Function Word Variability and the Scriptores Historiae Augustae." *Literary and Linguistic Computing*, vol. 13, no. 3, 1998, pp. 141–49, doi:10.1093/llc/13.3.141.

Tuccinardi, Enrico. "An Application of a Profile-Based Method for Authorship Verification: Investigating the Authenticity of Pliny the Younger's Letter to Trajan Concerning the Christians." *Digital Scholarship in the Humanities*, vol. 32, no. 2, 2017, pp. 435–47, doi:10.1093/llc/fqw001.

Tweedie, F. J., et al. "Neural network applications in stylometry: The Federalist Papers." *Computers and the Humanities*, vol. 30, no. 1, 1996, pp. 1–10, doi:10.1007/BF00054024.

Tweedie, Fiona J., and R. Harald Baayen. "How Variable May a Constant be? Measures of Lexical Richness in Perspective." *Computers and the Humanities*, vol. 32, no. 5, 1998, pp. 323–52, doi:10.1023/A:1001749303137.

Underwood, Ted. *DataMunging:rulesets at master · tedunderwood:DataMunging · GitHub*. 2015, github.com/tedunderwood/DataMunging/tree/master/rulesets.

---. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, 2019.

---. *Understanding Genre in a Collection of a Million Volumes, Interim Report*. figshare, 2014, figshare.com/articles/Understanding_Genre_in_a_Collection_of_a_Million_Volumes_Interim_Report

Underwood, Ted, et al. *Word Frequencies in English-language Literature, 1700-1922 (0.2) [Dataset]*. 2015, doi:10.13012/J8JW8BSJ.

Upton, Graham, and Ian Cook. "Loess." *A Dictionary of Statistics*, Oxford University Press, 2014.

Usher, Stephen, and Dietmar Najock. "A statistical study of authorship in the corpus lysiacum." *Computers and the Humanities*, vol. 16, no. 2, 1982, pp. 85–105, doi:10.1007/BF02259738.

Vinen, Richard. *The Long '68: Radical Protest and its Enemies.* Penguin, 2019.

Viola, Lorella, and Jaap Verheul. "Mining ethnicity: Discourse-driven topic modelling of immigrant discourses in the USA, 1898-1920." *Digital Scholarship in the Humanities*, vol. 35, no. 4, 2019, pp. 921–43, doi:10.1093/llc/fqz068/5601610.

Waldron, Levi, et al. "Optimized application of penalized regression methods to diverse genomic data." *Bioinformatics*, vol. 27, no. 24, 2011, pp. 3399–406, doi:10.1093/bioinformatics/btr591.

Wernli, Didier, and Frédéric Darbellay. *Interdisciplinarity and the 21st century research-intensive university.* LERU Universities; LERU Universities, 2016, www.leru.org/files/Interdisciplinarity-and-the-21st-Century-Research-Intensive-University-Full-paper.pdf.

Williams, Raymond. *Culture and Society 1780 - 1950.* Vintage, 2017.

---. *Modern Tragedy.* Verso, 1979.

---. *The Politics of Modernism: Against the New Conformists.* Verso, 2007.

---. "When Was Modernism?" *New Left Review*, vol. 1, no. 175, 1989, pp. 48–52, newleftreview.org/I/175/raymond-williams-when-was-modernism.

Wilson, Edmund. *Axel's Castle: A Study in the Imaginative Literature of 1870 - 1930.* Farrar, Straus, Giroux, 2004.

Wittig, Susan. "The computer and the concept of text." *Computers and the Humanities*, vol. 11, no. 4, 1977, pp. 211–15, doi:10.1007/BF02396857.

Wordsworth, William, and Samuel Taylor Coleridge. *Lyrical Ballads 1798.* Project Gutenberg, 2011, www.gutenberg.org/files/9622/9622-h/9622-h.htm.

Zanin, M., et al. "Combining complex networks and data mining: Why and how." *Physics Reports*, vol. 635, Elsevier B.V., 2016, pp. 1–44, doi:10.1016/j.physrep.2016.04.005.