

Close Reading with Computers

Textual Scholarship, Computational Formalism, and David Mitchell's *Cloud Atlas*

Brief Synopsis

This book is the first full-length monograph to bring a range of computational methods to bear on literature at the micro-level. While most contemporary digital studies are interested in distant-reading paradigms for large-scale literary history – using their digital methods as a telescope – *Close Reading with Computers* instead asks what happens when such techniques function as a *microscope*.

Focusing on the popular, award-winning, and genre-bending novel by David Mitchell, *Cloud Atlas* (2004), the book covers a range of digital approaches to close reading this text.¹ After a broad introduction to digital approaches and literature, the book begins with a textual-genetic publishing history of the novel and a range of visualization techniques that can describe and explicate the publishing and editorial processes. The second chapter then moves to an analysis of stylometric techniques for appraising micro-tectonic linguistic shifts in generic register. In particular, this chapter appraises the underlying assumptions of many authorship attribution techniques and reconfigures this approach for genre detection. The third chapter focuses on etymological mimesis and contemporary re-imagination of historical style. This is achieved through a set of digital etymological dating techniques and corpus frequency analyses. The book concludes with a section that more thoroughly interrogates the evidential findings of the computational work.

The book is a literary studies research monograph but also has an exemplary function; demonstrating a set of methods, accompanied by open-source software tools that I have developed, that others can use in their own literary-critical practices. In this way, I hope that the volume bridges the chasm between those using digital methods and those engaged in more traditional literary critical endeavours.

Biography

Martin Paul Eve is Professor of Literature, Technology and Publishing at Birkbeck, University of London. Martin holds a Ph.D. from the University of Sussex and is the author of four other books: *Pynchon and Philosophy: Wittgenstein, Foucault and Adorno* (Palgrave, 2014); *Open Access and the Humanities: Contexts, Controversies and the Future* (Cambridge University Press, 2014); *Password* (Bloomsbury, 2016); and *Literature Against Criticism: University English and Contemporary Fiction in Conflict* (Open Book Publishers, 2016). He has held three large grants

¹ *Cloud Atlas* won the British Book Awards Literary Fiction Award and the Richard & Judy Book of the Year award. The novel was short-listed for the 2004 Booker Prize, the Nebula Award, and the Arthur C. Clarke Award. A film of the book achieved widespread recognition in 2012.

from the Andrew W. Mellon Foundation and before beginning his academic career, Martin worked as a computer programmer.

Past Endorsements

“Martin Paul Eve is one of the most brilliant scholars of his generation. His ground-breaking *Literature Against Criticism* combines new and insightful readings of contemporary novelists (from Jennifer Egan to Tom McCarthy and from Sarah Waters to Percival Everett) who are in animated competition with university English. There are very few authors who can combine ethical, political and aesthetic readings of the contemporary novel with an encyclopaedic knowledge of the modern university.” – Bryan Cheyette, Professor of English, University of Reading

“Conjuring our passwords has become a daily act of our computer-saturated existence. By no means sequestered to our digital present, Martin Paul Eve's excellent account of the password covers its long and lively history. Weaving literary references with lucid technical explanations, Eve skillfully traces the evolution of password to probe its fundamental connections to issues of human identity, trust, and ownership.” – Gabriella Coleman, Wolfe Chair in Scientific and Technological Literacy, McGill University, Canada

"Martin Paul Eve's *Pynchon and Philosophy* is a work of consummate scholarship. Breaking new ground in Pynchon studies, Eve offers an immensely erudite, detailed and in-depth account of the ways in which the ideas of Wittgenstein, Foucault and Adorno help us to think about his texts. A first-rate book." - David Cowart, Louise Fry Scudder Professor of Humanities, University of South Carolina, USA

Reader Responses

Some portions of the book have appeared or are forthcoming in article form.

Excerpted peer review report from *SubStance* journal (University of Wisconsin Press):

"This article is seductive. First, it discusses the broad lines of possible arguments about the place of digital analysis in literary criticism. Then it presents stylometry as a methodology. Then it discusses the underlying suppositions about style and authorship that go unexpressed in too many digital analyses. And only then does it present some samples of analysis and what they might mean in Mitchell. I have rarely seen a digital analysis of style that was more clearly presented."

"So far, I haven't been impressed by digital stylometry, mainly because it seems to ignore the very philosophical and historical assumptions it was carrying. But this paper is a pleasure to read for its conscientious exposition of the assumptions of stylometry and its very excellent examples."

An excerpt of Chapter Two was published in the *Open Library of Humanities* and was covered by the media on the front page, online, of the *Guardian* newspaper and in the *De Standaard* national press in Belgium.

Proposed Table of Contents

I: Introduction

II: The Contemporary History of the Book

III: Reading Genre Computationally

IV: Historical Fiction and Linguistic Mimesis

V: Interpretation

Appendix A: Textual Variants of Cloud Atlas

Appendix B: Etymological Word Datings of Ewing Part I

Bibliography

Projected extent: 70,000 words.

Chapter-by-Chapter Summary

I: Introduction [10,000 words]

This initial chapter sets the scene for a close study of literature with computers. Outlining the fierce debates and controversies in literary studies around the so-called “digital humanities” – including the claims for their neoliberalism, their utilitarianism, and a more-general assertion of their anti-humanism – the introduction then turns back to an examination of what computation actually does: an endlessly patient, iterative, quantitative set of operations upon an incoming dataset.

From this observation, I then discuss the developments in large-scale literary history that have been facilitated by such work. Citing figures such as Ted Underwood, Matt Jockers, and Franco Moretti, I explore the ways in which computation’s labour-saving power has, by and large, been seen as a way to explore literary history and to overcome the limitations in reading labour time that are posing difficulties in other spaces (see, for instance, Amy Hungerford’s concept of “critical not-reading”).

I then move to contest this view of what computation can bring to the study of literature. While most distant methods have to work closely with their literary source material as a type of “tuning exercise” (for instance Jockers’s sentiment analysis work on James Joyce) and while the pages of the journal *Linguistic and Literary Computing/Digital Scholarship in the Humanities* have seen several articles that train a close focus on literary texts, I here propose that such methods can be used as both literary-theoretical stimuli and as evidence-gathering techniques.

The metaphor through which I propose this shift to close focus is that of a microscope vs. a telescope. While the distant-reading perspective aims to gain a large-scale view of literary history while losing some optical resolution, a microscope performs the same task of amplification but enlarges features that are invisible to the naked eye.

I also argue that computational methods bring theoretical discussions and challenges back to the space of literary criticism. For instance, when we talk of stylometry (the measurement of style), what is meant by this? Many hundreds of vectors could combine to yield a definition of “style”, but without asking what computers are measuring, to assert this as a stylistic measure feels far fetched.

Finally, this chapter introduces the main literary focus of the book: David Mitchell’s *Cloud Atlas*. This 2004 novel won the British Book Awards Literary Fiction Award and the Richard & Judy Book of the Year award. The novel was short-listed for the 2004 Booker Prize, the Nebula Award, and the Arthur C. Clarke Award. Indeed, the book is a highly significant work of contemporary literary fiction in its own right. However, the novel also possesses characteristics that make for a worthwhile computational study. Written in six wildly different linguistic and generic registers, with substantial version variance across editions in the US and the UK, *Cloud Atlas* presents itself as a case study for the questions of authorship, textual production, and historical fiction that I cover in this work.

This closing aspect also has a section on legal approaches to the computational study of contemporary fiction. In order to obtain a text for digital processing, I re-keyed the original novel. In these final words I explain the complex legal situation for digital work on contemporary novels in the era of Digital Rights Management.

II: The Contemporary History of the Book [15,000 words]

In his “Contemporary Fiction: Towards a Manifesto”, published in *Textual Practice*, Robert Eaglestone laments that the sites of publishing labour “do not see the point of” academics. He further points out that, despite academia being entrenched in the market economies of book circulation, one understudied aspect of “contemporary fiction is what we might call the ‘contemporary history of the book’: the ways in which the business of publishing helps to shape and control contemporary fiction. There seems to be a dearth of research into this aspect of the field.”

In this chapter I turn my attention to the publishing history and version variants of David Mitchell’s *Cloud Atlas*. For, in 2003, David Mitchell’s editorial contact at the US branch of Random House moved from the publisher, leaving the American edition of *Cloud Atlas* without an editor for approximately three months. Meanwhile, the UK edition of the manuscript was undergoing a series of editorial changes and rewrites that were never synchronised back into the US edition of the text. When the process was resumed at Random House under the editorial guidance of David Ebershoff, changes from New York were likewise not imported back into the UK edition. In the section entitled “An Orison of Sonmi ~451” these desynchronised rewritings are nearly total at the level of linguistic expression between UK and US paperbacks/electronic editions and there are a range of sub-episodes that only feature in one or other of the published editions.

This chapter consists of three parts. The first examines this dearth of textual scholarship in the space of contemporary fiction, with reference to other contemporary work that circles around this question, such as Matthew Kirschenbaum’s recent *Track Changes: A Literary History of Word Processing*. The second traces the history of the manuscript and typescripts, commenting upon the

degree of textual variance. The final section uses a set of computational visualizations to 1.) plot the stemma of the particular manuscript versions and their transmission histories (for instance, through translation and film adaptation) and 2.) to conduct a hermeneutic exercise in visualizing Ebershoff's line edits to the "Orison of Sonmi ~451" section.

The aims for this chapter are to show that version variance in contemporary fiction is a phenomenon worth studying in its own right; to unpick the actual literary production history of Mitchell's novel; and to demonstrate a range of computational methods that can be used to interpret line edits between manuscripts and to understand transmission histories of texts.

III: Reading Genre Computationally [15,000 words]

This chapter considers what it means to appraise genre with the aid of computers. For David Mitchell's *Cloud Atlas* contains six different and distinct generic registers. The first segment here explores the assumptions and limitations of computational stylometry. These are grouped around three core assumptions: 1. that there is a 'stylistic naturalism' of an author; 2. that stylometry measures subconsciously inscribed features of a text; 3. that authorship is the underlying textual feature that can be ascertained by the study of quantified formal aesthetics.

Having explored the literary-critical implications of these assumptions, the second part of this chapter explores computationally the linguistic mechanisms that create Mitchell's genre effects. Specifically, I show that authorship attribution techniques incorrectly cluster the chapters of *Cloud Atlas* as distinct "authors" using anything above the nineteen most-common words. This has implications for understandings of literary style and authorship since there is a consensus from what is usually used as an authorship attribution algorithm that Mitchell's generic segments are written by different authors; an adversarial stylometric approach.

The next part of this chapter compares the seafaring parts of Mitchell's novel with the writings of Herman Melville using Burrows's delta method. Despite the fact that the text claims, explicitly, that Melville is a point of emulation – noting that "Ewing puts me in mind of Melville's bumbler Cpt. Delano in 'Benito Cereno'" – the methods of authorship attribution do not link Melville and the Ewing chapter. This is the first step towards a broader claim: that Mitchell's episodes possess enough generic distinction to separate them from one another, *as though they were written by different authors*. In other words, this section both demonstrates one claim while disproving another. Certainly, Mitchell and Melville can be told apart using computational methods (the claim that Mitchell's writing imitates Melville is false for the computational approach). However, Mitchell's sections are also deemed sufficiently different here as to render them equally as distinct from one another as Melville is from Mitchell. That is, Mitchell does not emit a "Melville signal" (while *Moby-Dick* and "Benito Cereno" do) but he also does not emit a coherent "Mitchell signal".

The final section of this chapter consists of a part-of-speech trigram visualization and analysis. This is a way of making visually clear specific part-of-speech formulations that only appear in one section or another with a specific degree of additional frequency (e.g. 1.9 standard deviations). This shows, in the first place, ways in which visualizations can be a finding aid for particular linguistic

claims, but here also reveals the unique present-tense linguistic phrasings (NNP NNP VBZ and NNP VBZ DT) that lend pace to the Luisa Rey section of the novel.

Overall, this chapter is designed to demonstrate the ways in which we can understand the micro-tectonic linguistic shifts in novels, through a set of computational methods. It asks fundamental questions of what we mean by “literary style” and by what measures we can group different forms of writing. It uses visualization as an actual finding tool for points of interest, to then be re-synthesized into close readings. In short, this chapter demonstrates the principles of the digital microscope that I outlined above.

IV: Historical Fiction and Linguistic Mimesis [15,000 words]

What does it mean to write as though in some bygone period? This chapter addresses issues of mimetic accuracy in historical fiction that purports to come from a particular timeframe.

Specifically, the first section of *Cloud Atlas* purports to be set in 1850. Narrative clues approximately date the intra-diegetic diary object of this chapter between the period 1850 to 1910. This chapter argues, though, for the construction of a stylistic historical imaginary of this period's language that is not based on mimetic etymological accuracy. Specifically, I show that of the 13,246 words in The Pacific Diary of Adam Ewing, there are at least three terms that have an etymological first-usage date from after 1910: spillage, from ~1934; latino, from ~1946; and lazy-eye, from ~1960. I also show that racist and colonial terms occur with much greater frequency in *Cloud Atlas* than in a broader contemporary textual corpus (the Oxford English Corpus), indicating that the construction of imagined historical style likely rests more on infrequent word use and thematic terms from outmoded discourses than on etymological mimesis. Finally, I argue that there are political implications to the “puncturing” of linguistic accuracy for our consideration of Ewing and its colonial rhetorics.

The computational approaches to word dating and corpus linguistics in this chapter are framed around conventional discussions of historical fiction and mimetic accuracy. Specifically, I reference Jerome de Groot on the “heft and authenticity” that one is apparently supposed to convey in a work of historical fiction and ask what a divergence from such “authenticity” does to *Cloud Atlas*. I also here turn to aspects of parody and pastiche, questioning whether the puncturing of realism here constitutes another violation of the text's diegetic layers, as with its other metafictional tricks.

V: Interpretation [15,000 words]

This final chapter takes the preceding computational analyses and synthesises the results into a close reading of *Cloud Atlas* that focuses on the idea of the “archive” as central to the novel.

In terms of the second chapter's analysis of textual variance, I note that the chapter with the most obvious alterations is the part of the narrative where an archivist attempts to store the voice of a death-penalty convict for posterity. Yet, in the novel's trans-textual variance, these editions are different, thereby undoing the preservation effect that such an archive is supposed to have.

This argument then moves from the first chapter to the second, arguing that the disaggregation of subjectivity in the novel is furthered by the direct fracturing of an authorial coherence. Indeed, through the specific stylistic measures that I outline in chapter three I argue that a fractured and fragmented voice narrates the novel.

Finally, I argue, this extends even further into a fragmentation and puncturing of history. While other critics have discussed the work of time and history in this novel – Caroline Edwards, Courtney Hopf, Patrick O'Donnell, and many more – I argue for a most specific placement of the reader at the end of time. For, the mimetic cracks in the language of the Pacific Diary of Adam Ewing call for a re-interrogation of the “authenticity” of various performed language contexts. If one dating system – that in *Letters from Zedelghem* – cannot be trusted, and if the archive of both authorial self, and preservation of characters are also fragmented, then the reader-as-literary-historian is placed in a new relation to the text.

V: Appendices

Two appendices provide the underlying data on which the work has been conducted.

Estimated Delivery Date

The underlying computational work is now complete and the manuscript is coming together. I anticipate delivery before July 2018.

Figures and Tables

The manuscript will have approximately 15 figures that can be customized for black-and-white printing. There are also several tables (approximately 6 at present, including one lengthy appendix table).

Fit with Existing Literature

This work performs a different function to other digital humanities (DH) monographs on distant reading. While, for instance, Franco Moretti's *Distant Reading* and *Graphs, Maps, and Trees* or Matt Jockers's *Macroanalysis* back out to the larger scale, this work provides a range of computational approaches for the *close reading* and *interpretation* of literary fiction. In this sense it attempts to reconcile a digital form of practice with conventional close reading methods.

The work is also aligned with, but differentiated from, the forms of computational poetics and thinking, such as that recently outlined by Dennis Tenen in his *Plain Text: The Poetics of Computation*. In one sense, the work does take much of the theoretical thinking from the social DH and New Media spaces (Alan Liu, Stephen Ramsey, N. Katherine Hayles, Susan Schreibman and others).

In another sense, though, this book is also a literary studies monograph on the work of David Mitchell. There is a burgeoning critical scene around Mitchell's work. Patrick O'Donnell's *A Temporary Future* was the first monograph on this author, while Rose Harris-Birtill has a second

under submission. There has been a special issue of *SubStance* journal on Mitchell's work and an edited collection compiled by Sarah Dillon of Cambridge University.

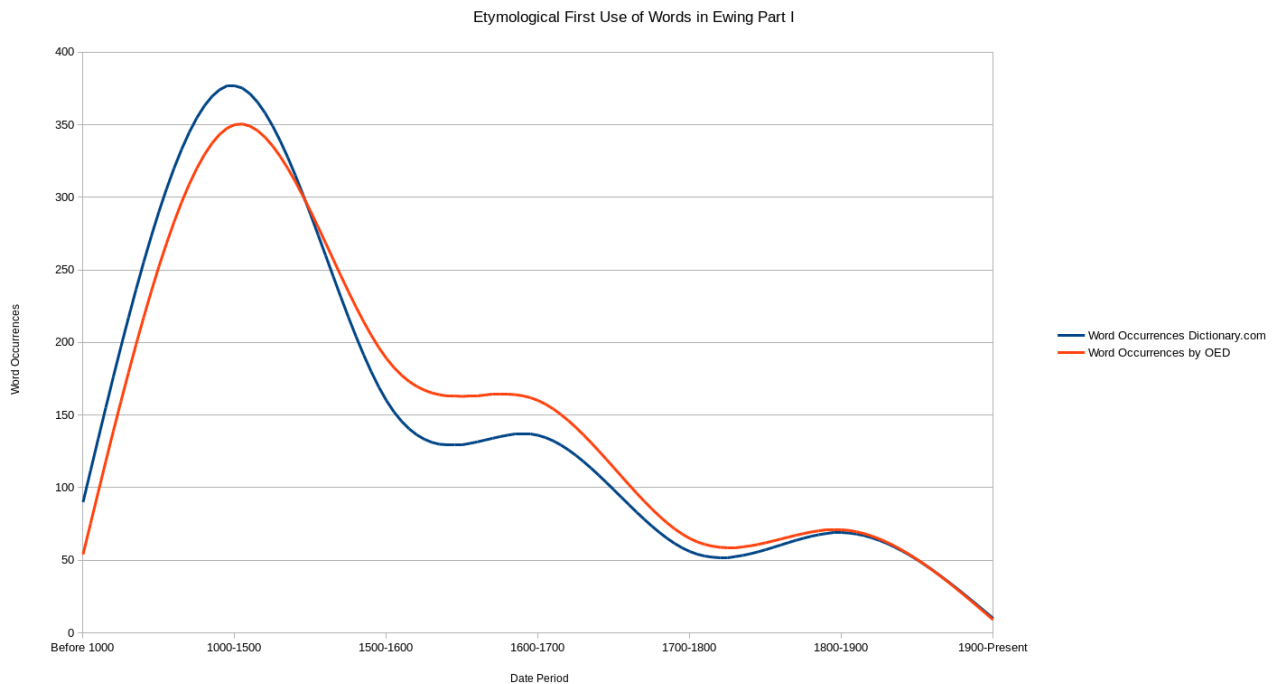
Thus, while the single-author – and even single-text – approach is unorthodox, it is both merited on scholarly interest but also on thematic coherence. If there is to be an approach to close reading, then, I would argue, this focus should truly be *close*.

Intended Audience

The intended audiences for this book are:

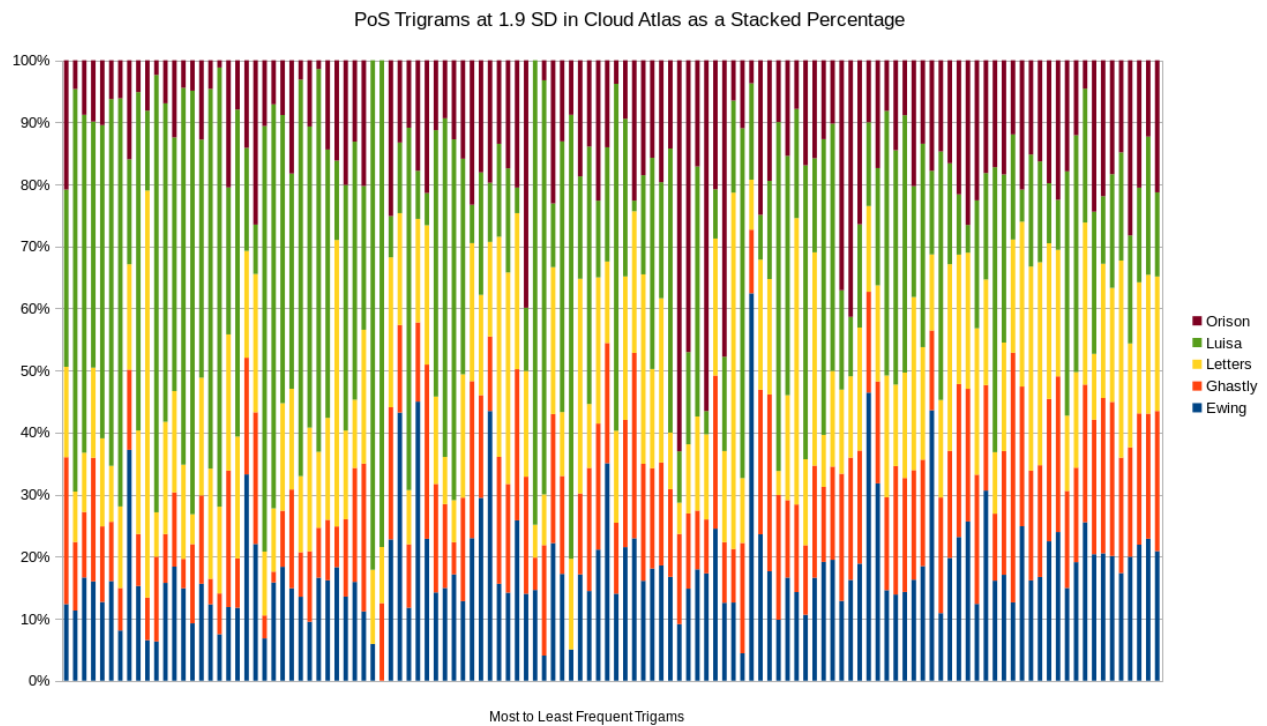
- Literary scholars interested in digital methods for close reading
- Contemporary literary scholars interested in textual scholarship and version variance
- Literary scholars interested in the work of David Mitchell
- Digital humanists interested in approaches for close reading
- Students of any of the above

Sample Visualizations



Sample 1: Etymological First-Usage Dates of Words in Ewing Part I (purported to be written in 1850)

This diagram is a profile of the etymological first-usage dates of terms in Part I of the Pacific Diary of Adam Ewing, which is purported to be written in 1850 in the novel. There are two etymological sources: the OED and dictionary.com. The diagram clearly shows a violation of a realist linguistic mimesis of the period.



Sample 2: Anomalous Part-of-Speech Trigram Visualizations for the Different Sections of *Cloud Atlas* at 1.9 Standard Deviations. This shows a means of navigating uniquely more or less frequent part-of-speech formulations across genres.

This diagram provides a way to identify part of speech formulations that occur significantly more frequently in one section of the novel than another. In chapter three I set out a range of distinctive voicings that can be identified by this approach and tie them to specific genre effects.