Close Reading with Computers: Genre Signals, Parts of Speech, and David Mitchell's *Cloud Atlas*

Martin Paul Eve

Close Reading with Computers: Genre Signals, Parts of Speech, and David Mitchell's *Cloud Atlas* ᵭ

Martin Paul Eve

Close Reading, Distant Reading, and Labor

Reading literature with the aid of computational techniques is controversial. For some, digital approaches apparently fetishize the curation of textual archives, lack interpretative rigor (or even just interpretation), and are thoroughly 'neoliberal' in their pursuit of Silicon Valley-esque software-tool production (Allington, Brouillette, and Golumbia; see "Editors' Choice" for a good range of counter-responses). For others, the potential benefits of amplifying reading-labor-power through non-consumptive use of book corpora fulfills the dreams of early twentieth-century Russian formalism and yields new, distant ways in which we can consider textual pattern-making (Jockers; Moretti, *Distant Reading*; Moretti, *Graphs*). Indeed, there are many arguments to be made around the quantifying processes of computational stylometry that the humanities are – and should be – qualitative in their approaches. At the same time, we also know that the humanities do not hold a monopoly on aesthetics; mathematics, statistics, and computation have a beauty and intuition behind them that are as human as any works of art and need not demean the aesthetics of objects with which they have contact.

Among the best metaphors that we might use for computational methods in literary studies is that of a telescope, allowing us, at a distance, to ingest, process, and perhaps understand texts within grand perspectives, even while losing some detail of the image. Literary history, we are told, can be seen unfolding over vast time periods when we simply do not have the time in our lives to read that many novels (Moretti, "The Slaughterhouse"). This allows, for instance, for the large-scale mappings of genre formations and their lifecycles over time (Underwood). In each of these cases, the computer becomes the tool that can read on our behalf; we will delegate reading labor to the machine and then expend our

interpretative efforts upon the resultant quantitative dataset. For, as Lisa Gitelman and others have rightly told us: there is no such thing as 'raw data' (Gitelman).

Yet, the computer can also act as a microscope. While both the telescope and the microscope have powers of amplification, the question becomes: what can the computer see, in its repetitive and unwavering attention to minute detail, that is less (or even in-) visible to human readers? This question has occurred to others, although it is a less common way of operating, and I do not propose it as a novelty even while I aim here to invite a broader audience to the table. For instance, the esteemed journal *Literary and Linguistic Computing* (recently renamed *Digital Scholarship in the Humanities*) has featured, over the past three years, two papers that examine single texts in detail, working on authorship attribution fingerprints (Pearl, Lu, and Haghighi; Gladwin, Lavin, and Look).

Somewhere between these micro and telescopic scales sits David Mitchell's *Cloud Atlas* (2004), the text to which I turn my focus in this article. This novel, divided as it is into six generically distinct registers, with a pyramid-style cascade towards the future in which each section breaks halfway only to move to the next, deals with a vast and telescopic history. Casey Shoop and Dermot Ryan, for instance, locate the novel within the space of 'Big History' (Shoop and Ryan 101). On the other hand, almost every critic of the novel has remarked upon the linguistic play of the text and Mitchell's seeming protean ability to shift between genre styles at will (see, for just a small collection, O'Donnell; Dimovitz; Hopf). Critics have also noted the novel's incursion into the digital space, with its imitations of new media ecologies that John Shanahan has called the text's "digital transcendentalism" (Shanahan). It was, then, the way in which *Cloud Atlas* mediates a colossal philosophical historiography through minute and detailed attention to linguistic morphology within a new media frame that attracted me to use the novel as a study of what might be possible for digital close reading and that I here present. For *Cloud Atlas* seems to effect the very compression of reading labor time that is desired from computational approaches to big literary history through its language games.

If distant-reading techniques are supposed, though, to save reading labor, then it is an irony that using a literary-computational microscope to study a contemporary novel such as *Cloud Atlas* remains a great deal of work. For, in the UK where I live and work, as of 2017, there is a provision in law that implements EU Directive 2001/29/EC. This dry directive states that it is a criminal offence to break the DRM on digital files. In other words, it is illegal, even for personal or research purposes, to remove the DRM from a purchased Amazon Kindle file. There are supposed to be

protections in the directive to allow personal use or research upon such texts. Indeed, the act states that:

> Notwithstanding the legal protection provided for in paragraph 1, in the absence of voluntary measures taken by rightsholders, including agreements between rightsholders and other parties concerned, Member States shall take appropriate measures to ensure that rightsholders make available to the beneficiary of an exception or limitation provided for in national law in accordance with Article 5(2)(a), (2)(c), (2)(d), (2)(e), (3)(a), (3)(b) or (3)(e) the means of benefiting from that exception or limitation, to the extent necessary to benefit from that exception or limitation and where that beneficiary has legal access to the protected work or subject-matter concerned.

In the UK, this is implemented in Section S296ZE of the Copyright, Designs and Patents Act. Section S296ZE provides a way to contest situations wherein the rightsholder's Technological Protection Measures prevent an authorized exempted use, thereby implementing the EU directive. This involves a twofold process of: 1. asking a publisher to voluntarily provide a copy that can be used in such a way; 2. contacting the Secretary of State to ask for a directive to yield a way of benefiting from the exemption on Kindle format books for non-commercial academic research purposes. This process is known to be both very time-consuming and to have little chance of providing the desired exemption.

In order to remain within the bounds of the law, I opted to manually retype the text from the Kindle (or E-edition) of the novel (for information on the version variants of the novel, see Eve, "'You Have to Keep Track of Your Changes': The Version Variants and Publishing History of David Mitchell's *Cloud Atlas*"). This was both a tiring and tiresome endeavor and I hope that at some point in an enlightened future, digital versions of copyrighted works of fiction might be available to purchase in forms that will allow computational research to be conducted upon them. For now, though, suffice it to say that it remains an incredibly labor-intensive process even to get to the point where one has a research object upon which it is possible to work. This is why I refer, though, to the techniques that I here conduct through the analogy of a microscope, rather than any kind of 'distant reading.' For it has saved me *no reading labor* using computational methods to study a single contemporary text that is under copyright. Indeed, in retyping the novel, I have read the text more closely than I have ever previously read any other literary or critical work. Yet, without the computational methods, I still *could not see*. The computational micro-, rather than macro-, scope can teach us things about texts that we could see with our own eyes were we infinitely patient and obsessive. But we are neither of these things.

Successes and Failures of Computational Stylometry

What does it mean to 'write like David Mitchell in *Cloud Atlas*'? One of the most basic things that we can do with computational techniques is to conduct an analysis of the most-frequently used words in a text. That doesn't sound very exciting on its own, but it turns out that the subconscious ways in which authors use seemingly insignificant words is an extremely effective marker for authorship attribution. That is, most texts by the same author can be accurately clustered by comparing the 'Manhattan distance' plots of the z-scores (that is, the standard deviation) of each word frequency within a work. I wondered, though, what would happen if I undertook such an analysis on each section of Mitchell's novel. Would the underlying – and presumed subconscious elements of language – change between sections? Or would we, in fact, end up with Mitchell's persona inscribed within these texts? A set of stylometric techniques can help us to answer some of these questions.

As the name implies, computational stylometry is the measurement ('metry') of stylistic properties of texts ('stylo') using computers. Stylometry, as a quantifying activity, has a long and varied history, from legal court cases where the accused was acquitted on the basis of stylometric evidence, such as that of Steve Raymond (or speculative/hypothetical legal approaches), to authorship attribution (see the widely discredited Morton 205-6; but also Juola, "Stylometry"). In the latter case, as charted by Anthony Kenny, the discipline dates back to approximately 1851 when Augustus de Morgan suggested that a dispute over the attribution of certain epistles could be settled by measuring average word lengths and correlating them with known writings of St Paul (Kenny 1). At the time of writing, it is claimed that computational forensic stylometry "can identify individuals in sets of 50 authors with better than 90% accuracy, and [can] even [be] scaled to more than 100,000 authors" (Stolerman et al. 186).

In terms of a background to stylometry, a significant breakthrough, or at least a 'key moment' of success, took place around 1964 with the publication of Mosteller and Wallace's work on the set of pseudonymously published *Federalist* papers of 1787-1788, which were pushing for the adoption of the proposed Constitution for the United States. Mosteller and Wallace analyzed the distribution of 30 function words throughout the *Federalist* papers and managed to come to the same conclusion of authorship as the historians, based on statistically inferred probabilities and Bayesian analysis (Mosteller and Wallace). As Juola frames it, there are several reasons why this corpus formed an important test-bed for stylometry:

> First, the documents themselves are widely available [...], including over the Internet through sources such as Project Gutenberg. Second, the candidate set for authorship is well-defined; the author of the disputed papers is known to be either Hamilton or Madison. Third, the undisputed papers provide excellent samples of undisputed text written by the same authors, at the same time, on the same topic, in the same genre, for publication via the same media.

In Juola's words, "[a] more representative training set would be hard to imagine" (Juola, "Authorship" 242–243).

If, though, the *Federalist* papers represent a significant success for stylometric authorship attribution, there have also been some disastrous failures. In the early 1990s, a series of criminal court cases turned to forensic stylometry to identify authorship of documents (for example, Thomas McCrossen's appeal in London in July 1991; the prosecution of Frank Beck in Leicester in 1992; the Dublin trial of Vincent Connell in December 1991; Nicky Kelly's pardon by the Irish government in April 1992; the case of Joseph Nelson-Wilson in London in 1992; and the Carl Bridgewater murder case) (Holmes 114; Juola, "Authorship" 243). Indeed, it is frequently the case that court trials turn upon the authorship of specific documents, be they suicide notes, sent emails, or written letters (Chaski, "Who's at the Keyboard"). These specific cases, however, all relied on a particular technique known as 'qsum' or 'cusum' – for 'cumulative sum' of the deviations from the mean – which is designed to measure the stability of a measured feature of a text (Farringdon). The only problem here was that, almost immediately, the cusum technique came under intense scrutiny and theoretical criticism, ending in a live televisually broadcast failure of an authorship attribution test using this method (Canter; Hardcastle, "Forensic Linguistics"; Hardcastle, "CUSUM"; Hilton; Holmes and Tweedie; Juola, "Authorship" 243–244). Despite this failing, specific stylometric techniques remain admissible as evidence in courts of law depending upon their credibility and the jurisdiction's specific laws on admissibility (Chaski, "The Keyboard Dilemma"; McMenamin; Juola, "Authorship" 307–316).

The other most well-known case of failure in the field of stylometry occurred in the late 1990s when Don Foster attributed the poem "A Funeral Elegy" to William Shakespeare using a raft of stylometric approaches (Grieve). The attendant press coverage landed this claim on the front page of the *New York Times* and the community of traditional Shakespeare scholars reacted in disbelief. That said, when Foster refused to accept traditional historicist arguments against his attribution, stylometric work by multiple groups of scholars pointed to John Ford as the far-more likely author of the poem, which Foster eventually accepted (Elliot and Valenza, "And Then There Were None"; Elliot and Valenza,

"The Professor Doth Protest"; Elliot and Valenza, "So Many Hardballs"). While, as Juola points out, "this cut-and-thrust debate can be regarded as a good (if somewhat bitter) result of the standard scholarly process of criticism," for many scholars it marked the only interaction that they have ever had with stylometry and the result could only be a perception of notoriety and inaccuracy (Juola, "Authorship" 245).

That said, there have also been, especially in recent years, some extremely successful algorithmic developments for detecting authorship. Perhaps the most well known of these is the 1992 so-called 'Burrows's delta' (Burrows). With apologies for a brief mathematical deviation, Burrows's delta (the word here meaning the mathematical symbol for 'difference': Δ) consists of two steps to conduct a multivariate statistical authorship attribution. First of all, one measures the most-frequent words that occur in a text and then relativizes these using a 'z-score' measure. A z-score measurement is basically asking: 'by how much does a word's frequency differ from the average deviation of the other words?' So, the first thing that we would calculate here is the 'standard deviation' of the entire word set. A standard deviation means the square root of the average of the squared deviations of the values from the average. Or, in other words: work out the average frequency with which words occur in a text; then work out (for each word) how many more or less times that word occurs relative to the average; then square this and add up all such deviations; then divide this by the number of words; then square root the result. To get the *z-score*, we next take an individual word's frequency, subtract the average (mean) frequency, and divide this result by the standard deviation of the whole set. This is conventionally written as score (X) minus mean (mu / μ) divided by sigma (standard deviation / σ):

$$\frac{X - \mu}{\sigma}$$

Once we have a ranked series of z-scores for each term, the second operation in Burrows's delta is to calculate the difference between the words in both texts. This means taking the z-score of, say, the word 'the' in text A and subtracting the z-score of the word 'the' in text B. Once we have done this for every word that we wish to take into account, we add all of these differences together, a move that is the mathematical equivalent of taking the 'Manhattan distance' (named because it moves in right angled blocks like the city of Manhattan, rather than going 'as the crow flies') between the multi-dimensional space plots of these terms (Argamon). In Burrows's delta, the smaller this total addition of differences is, the more likely it is that two texts were written by the same author.

Burrows's delta has been seen as a successful algorithm for many years, as validated in several studies (Hoover; Rybicki and Eder). It is, mathematically speaking, relatively easy to calculate and seems to produce good results. However, it is not entirely known *why* the delta method is so good at clustering texts written by the same author, although recent work has suggested that such a "text distance measure is particularly successful in authorship attribution if emphasizing structural differences of author style profiles without being too much influenced by actual amplitudes," as does Burrows's delta (Evert et al.).

Yet, Burrows himself was always cautious about what he was doing. When writing of 'authorial fingerprints,' for example, Burrows noted that "we do not yet have either proof or promise" of the "very existence" of such a phenomenon (Burrows 268). Burrows also points out that, "[n]ot unexpectedly," his method "works least well with texts of a genre uncharacteristic of their author and, in one case, with texts far separated in time across a long literary career" (Burrows 267). This brings us to a point where it is worth delving deeper into the underlying assumptions of many stylometric methods.

Assumptions about Writing Style

There are a number of supposed premises on which most stylometric studies rest and these pertain to its use as a means of identifying authorship. Before moving to work on *Cloud Atlas* it is worth briefly covering these since they bear more broadly on how we conceive of literary style. These assumptions are: 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.

The first of these assumptions, that there is a 'stylistic naturalism' to an author's works, is premised on the idea that most of us, when writing, do not consider how our works will be read by computers. As Brennan and Greenstadt put it, "in many historical matters, author-ship has been unintentionally lost to time and it can be assumed that the authors did not have the knowledge or inclination to attempt to hide their linguistic style. However, this may not be the case for modern authors who wish to hide their identity" (Brennan and Greenstadt). Language is a tool of communication between people, designed to convey or cause specific effects or affects. The stylistic features of texts are usually considered to be a contributor to the overarching impact of the communication. Indeed, the scansion and rhythm of a work of prose, for instance, is an important feature of well-written texts, the three-part list being a good example of this in persuasive works of rhetoric. Yet, at the same time, the selection

and prioritization of specific stylistic features (rhythm, cadence, word length, repetition) has knock-on effects to the other elements of language that are deployed.

In other words, and to put it bluntly: there are hundreds of stylistic traits of texts that we can measure and determine. It is not possible for an author to hold all of these in his or her working memory while writing and, instead, authors write for intended effects. The presumption that a reader will react in various ways to one's writing is, or at least should be, the overarching concern when writing. Yet, this leads to an idea of what I call a stylistic naturalism: the conceit that authors write in ways that are somehow blind to the processes of measurement of stylometry.

I would instead seek to re-couch this slightly differently. Any good author is aware that his or her writing is to be 'measured' – so to speak – by a reader. However, there is a constant play of balance at work here. In prioritizing one set of measurements – for instance, the long, rambling sentences of David Foster Wallace's *Infinite Jest* (1998) – others *must* be ignored. Authors are not unaware that they are being measured, they just must choose which measures are of most use for their literary purposes. This is a type of 'natural' writing then that can only be called natural in that it is social and not individual. Anticipated readerly reactions condition the writing process. As Patrick Juola puts it,

> the assumption of most researchers, then, is that people have a characteristic pattern of language use, a sort of 'authorial fingerprint' that can be detected in their writings. […] On the other hand, there are also good practical reasons to believe that such fingerprints may be very complex, certainly more complex than simple univariate statistics such as average word length or vocabulary size. ("Authorship" 239)

A sub-assumption that we might also put beneath the 'stylistic naturalism' claim is that authors behave in the same way when writing their various works; or, at least, that stylometric profiles do not substantially change even if authors deliberately try to alter their own styles. This also assumes that authors' own styles do not change naturally with time – a contentious claim (see the well-known Saïd). Indeed, in a 2014 chapter, Ariel Stolerman and colleagues identify shifting stylometric profiles of authors as a key failing in traditional "closed-world" settings (Stolerman et al.). (What Stolerman et al. mean by 'closed-world' here is that there is a known list of suspected authors and a computational classifier is trained to correctly attribute unknown works based on known stylometric profiles, rather than an environment where any author should be grouped apart from all others.) Yet, what happens, in stylometric terms, when an author such as Sarah Waters moves from a neo-Victorian mode to writing about the Second World War? What happens when Hilary Mantel writes about

Margaret Thatcher, as opposed to the Tudor setting of *Wolf Hall*? What happens when Sarah Hall moves from the feminist utopian genre of *The Carhullan Army* to the more naturalistic and contemporaneous setting of *The Wolf Border*?

These questions bring us to the obverse, but somehow linked counterpart, of the assumption that there might be a stylistic naturalism. That is, that stylometry can measure subconsciously inscribed elements of texts. As David I. Holmes puts it, at the heart of stylometry "lies an assumption that authors have an unconscious aspect to their style, an aspect which cannot consciously be manipulated but which possesses features which are quantifiable and which may be distinctive" (111). This is a different type of stylistic naturalism claim, one that, instead of asserting that authors are behaving in ways that make them unaware of stylometric profiling, looks instead to an author's subconscious as a site of unchangeable linguistic practice. Indeed, Freudian psychoanalysis has long held that aspects of communication and language harbor revelations about a person of which they have little or no control. Practical assaults against stylometric methods (known as adversarial attacks) have shown that, in such cases, some types of stylometry fare little better than chance against such methods (Brennan and Greenstadt 2). That said, as I will show shortly, all but one of the different narrative sections of *Cloud Atlas* can be distinguished from one another through the relative frequencies of the terms 'the', 'a', 'I', 'to', 'of', and 'in'. Yet, who among us, when writing, is conscious of the relative frequency with which we ourselves use these terms? These seemingly unimportant pronouns and prepositions are used *when we need them*, not usually as a conscious stylistic choice. In other words, the internalized stylistic profile of our individual communications usually determines how, why, and how frequently these terms are used; they are thought to be beyond our control. Such features are, therefore, conceived as subconsciously inscribed elements of a text that are difficult for an author to modify, even if he or she knows that stylometric profiling will be conducted upon a text. Yet, as I will go on to show, David Mitchell's novel, in its genre play, does manipulate such features.

All of which brings me to the final of the assumptions that I identify in most work on stylometry, namely that authorship is the underlying textual feature that can be ascertained by the study of quantified formal aesthetics. Of course, there are lengthy poststructuralist debates about what authorship actually means for the reception of texts (Barthes; Foucault; Burke). There are also disputes in labor and publishing studies as to how the individual work of 'authorship' is prioritized above all others, when actually there are many forms of labor without which publishing would not be possible: typesetting/text encoding, copyediting, proofread-

ing, programming, graphical design, format creation, digital preservation, platform maintenance, forward-migration of content, security design, marketing, social media promotion, implementation of semantic machine-readability, licensing and legal, and the list goes on (Eve, "Scarcity and Abundance"; Eve, "The Great Automatic Grammatizator"). So, the first challenge here for stylometry is to understand what impact these polyvalent labor practices have in the crafting of a single, authorial profile. We know, for example, that David Ebershoff requested substantial line edits to the US edition of *Cloud Atlas*. So, what sense does it then make to say that the figure identified as 'David Mitchell' would correlate to a stylistic profile of this text? At best, if the stylometry is working correctly as an attribution system centered on the author, it would identify this text as a harmonized fusion of Mitchell and Ebershoff.

The challenge that I actually want to pose to these three straw-figures that I have drawn up against many stylometric practices is one foreshadowed by Matt Jockers and others at the Stanford Literary Lab; namely that the author-signal is often neither the sole nor the most important signal that we can detect through stylometry (Jockers). Indeed, the first pamphlet of the Stanford Literary Lab found that, while the pull of the author-signal was strong and seemed even to outweigh other signals, various quantitative signatures also corresponded to those features that we might call 'genre' (Allison et al.). Instead, especially in the case of Mitchell's rich and varied novel, which was heavily edited by another person, and which deliberately employs mimicry and pastiche to achieve its proliferation of stylistic effects, it might be more appropriate to consider the *genre* signals that a text emits.

Understanding Mitchell's Genres Through Computational Formalism

In order to investigate the distinctions between the chapters of Mitchell's novels, the first thing that I was keen to check was whether the most basic methods of Burrows's delta analysis of z-scored Manhattan distances could correctly segment and group the different sections of *Cloud Atlas* within a hierarchical dendrogram. This would, I hoped, ascertain at the highest level whether Mitchell's writing is truly differentiated between chapters or whether there is an underlying authorial stylistic signature at work. Indeed, in a 2004 competition, the delta method met a good standard for competitive accuracy (Juola, "Authorship" 297). To do this, I used the 'stylo' package in R to ascertain the most frequent words (and then the most frequent bigrams for characters) in the whole novel and to hierarchically rank these and z-score them above the average for each section (R Core Team; Eder, Kestemont, and Rybicki).

Computing the Manhattan distance on each of these (for words and 2-character groupings) these rendered the following clusterings (Figures 1 and 2):
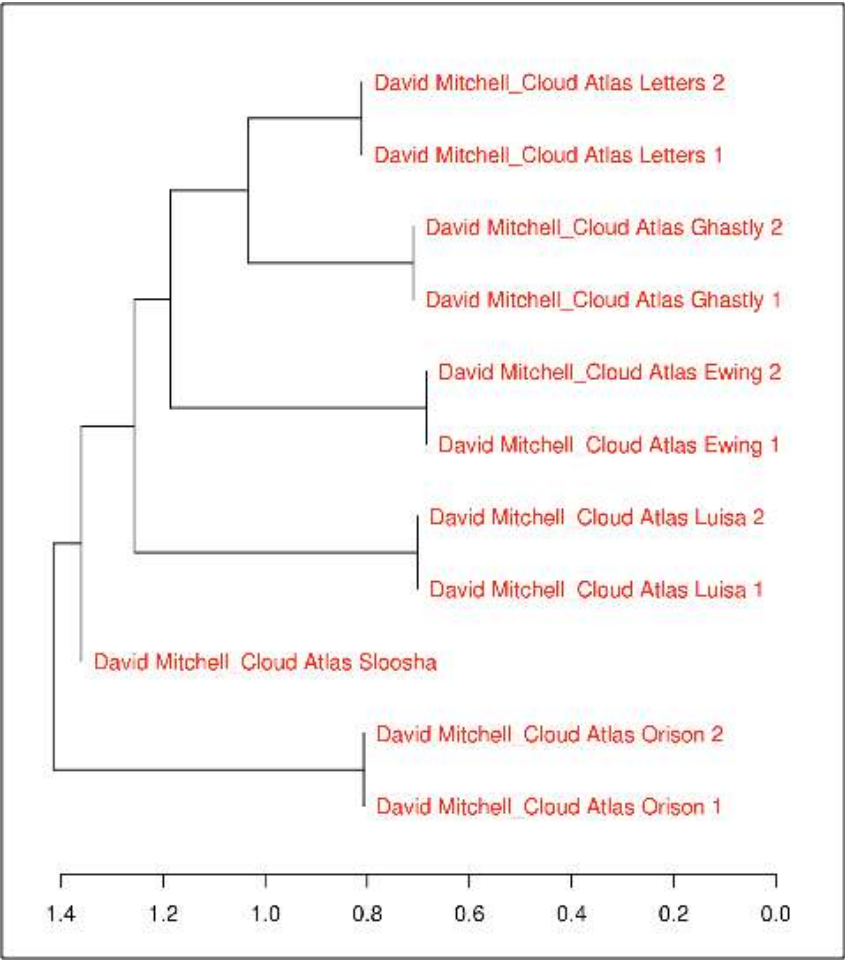

Figure 1: The sections of *Cloud Atlas* grouped by classic delta (z-scored 5,000 most-frequent-words differentiated by Manhattan distance).
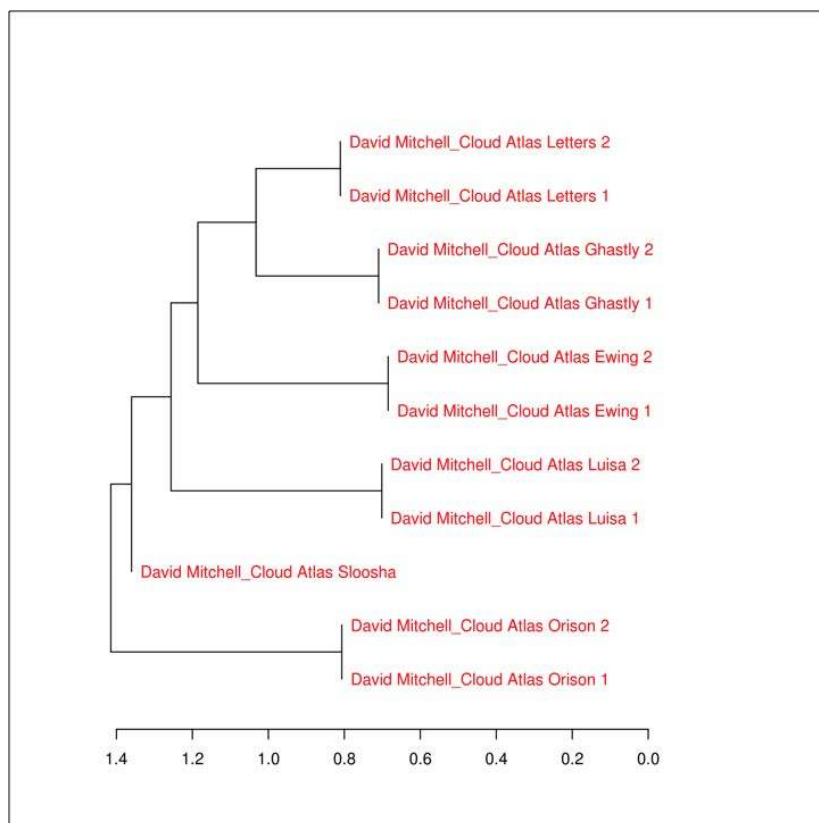
Figure 2: The sections of *Cloud Atlas* grouped by classic delta (z-scored 5,000 most-frequent-bigrams of characters differentiated by Manhattan distance).

What this shows us is not particularly sophisticated or novel, but it does verify the most cursory of stylometric phenomena here. Mitchell's novel is strongly differentiated between sections in terms of the unique lexical content and the order in which the most-frequent terms occur. This is the case whether we take the 5,000 most frequent words or the 5,000 most frequent bigrams. What is perhaps more curious is that the same holds true (although I haven't here pictured it) when one computes this based solely on words in the top 5,000 that occur in *all* of the narratives (of which there are 284, most of which are common words such as 'the'). In other words, the frequency with which Mitchell uses common words varies

enough between different sections of the text as to be able to statistically distinguish them from one another.

In fact, though, we can actually be far more granular than this in a description of the novel and its specific segments. With the exception of "An Orison of Sonmi ~451," the sections of *Cloud Atlas* can be distinguished from one another and grouped purely by how frequently Mitchell does or doesn't use the six most frequent words: 'the', 'a', 'I', 'to', 'of', and 'in'. When scored by the same classic delta paradigm as above, the only mistaken classifications are that "Orison Part I" is billed as part of "The Ghastly Ordeal of Timothy Cavendish" while "Orison Part II" is mistaken for a "Luisa Rey Mystery" segment. All other parts of the novel differ from each other by enough of a margin, but *only* in the use of these six words, as to make the chapters distinguishable from each other (Figure 3).

To accurately classify "An Orison of Sonmi ~451" with its counterpart requires an expansion to just the 20 most common words in the novel: 'the', 'a', 'I', 'to', 'of', 'in', 'and', 'my', 'was', 'you', 'an', 'it', 'his', 'for', 'me', 'but', 'on', 'that', 'he', and 'is'.

Such a low barrier of most-frequent-word counts as an accurate discriminator between the sections of Mitchell's novel is quite remarkable. However, the cluster dendrogram analysis method that I am using is hard to statistically validate. In other words, the question here is whether, if I ran this same procedure on other novels that did not share the stylistic variances of Mitchell's text, we might see random groupings, and what the statistical likelihood is that the groupings shown above have been arrived at by chance, rather than being distinct feature-sets of the sub-texts. After all, the fact that it was at the twenty-words mark that the clustering worked, and not below that, is arbitrary and based on my advanced knowledge of the dataset (the novel). This could lead to a type III error, or HARKing; hypothesizing after results are known (Kerr).

According to Maciej Eder, validation of cluster-analysis dendrograms can be undertaken, to an extent, by using a technique called bootstrap consensus tree plotting. Essentially, this technique re-runs the clustering algorithm over multiple iterations for many different most-frequent-word values and produces a final tree when a certain percentage of the underlying trees agree with each other. Running this same procedure on *Cloud Atlas* at 95% confidence, we would expect, from the above investigation, to see a correct clustering of all sections except for "An Orison of Sonmi ~451" (there are 284 shared words among all the sections and the cutoff point was 20 words, so the percentage of confidence here at which we would expect proper classification is: $100-((20/284)*100) = 92.9\%$). And, indeed, the following two diagrams (one at 95% and one at 92%) seem to give some validation to the findings (Figures 4 and 5).
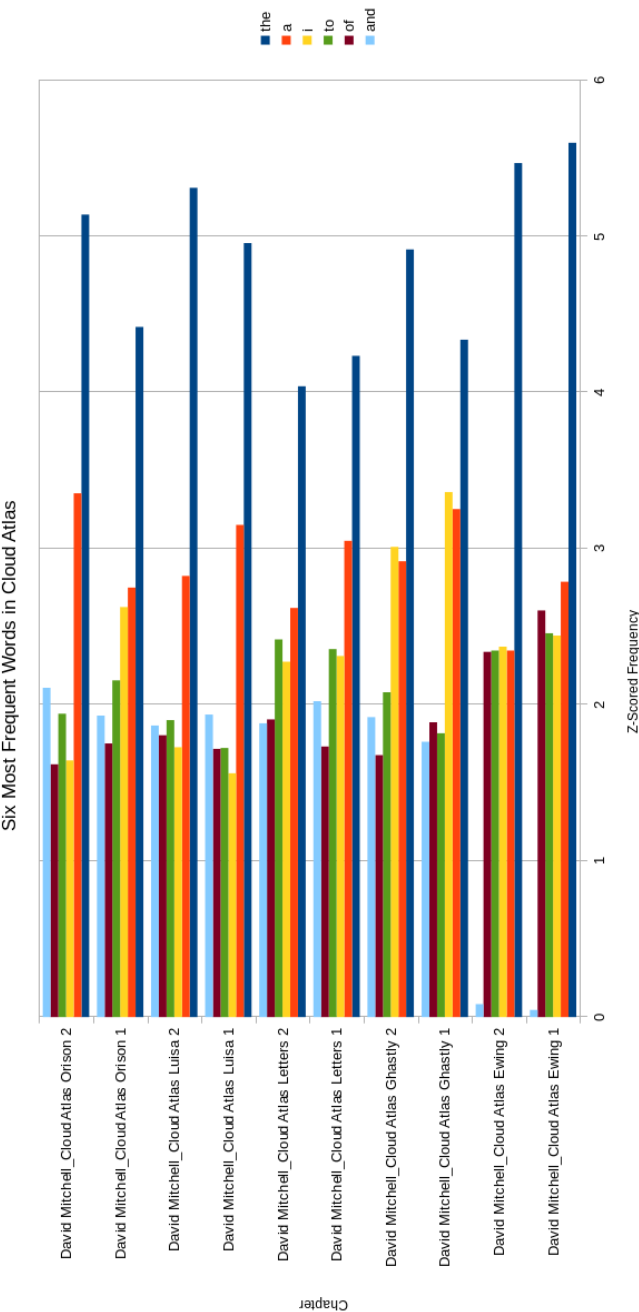
Figure 3: The *z*-scored frequency occurrence of the six most-frequent words in *Cloud Atlas* in all chapters except "Sloosha's Crossin'"

Figure 4: *Cloud Atlas E* classified using 1 to 284 most-commonly used and shared words in a bootstrap consensus tree with 95% consensus of underlying clusters. Note that all sections are clustered correctly except for "An Orison of Sonmi ~451," which is marked as a discrete section in each case.
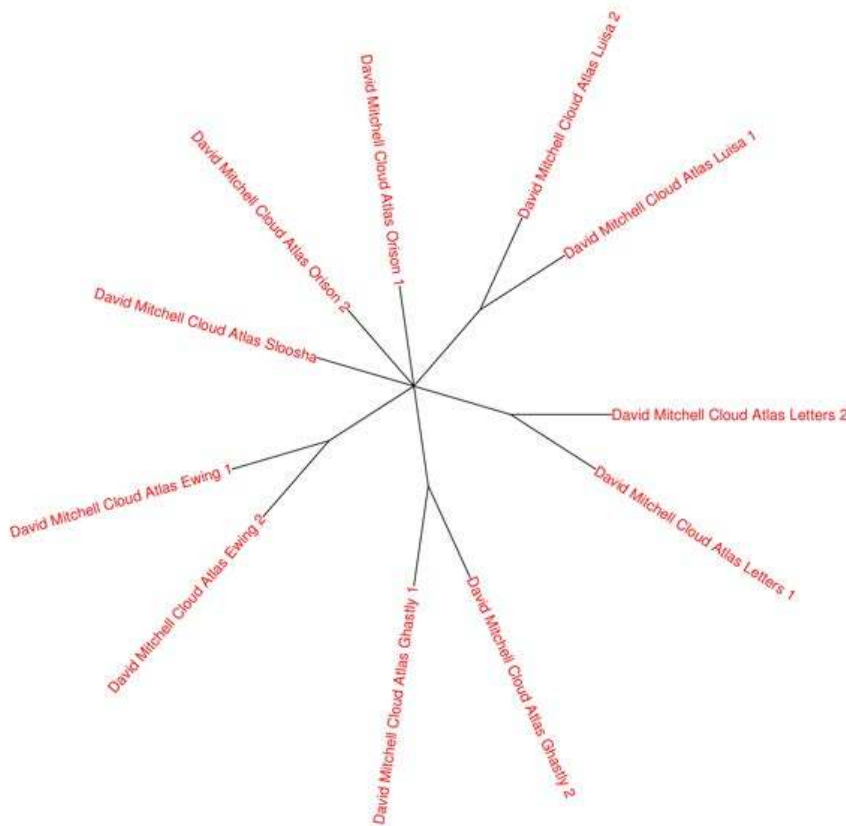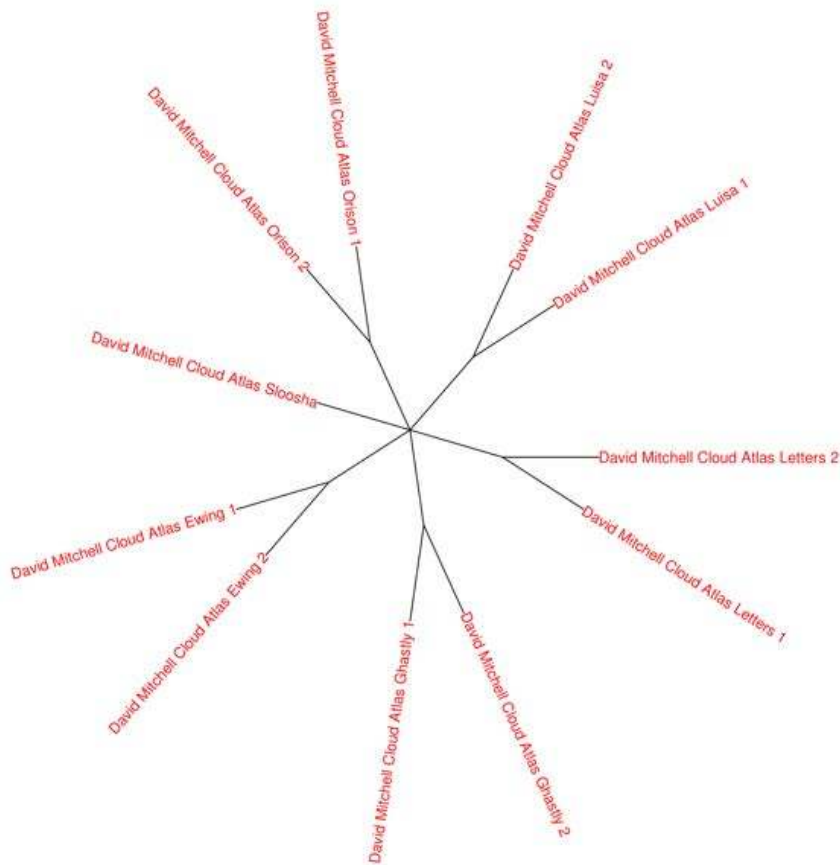
Figure 5: *Cloud Atlas E* classified using 1 to 284 most-commonly used and shared words in a bootstrap consensus tree with 92% consensus of underlying clusters. Note that here, as predicted, "An Orison of Sonmi ~451" is correctly classified.

This validation technique and underlying clustering analysis tells us a few things about the initial, internal stylistic properties of Mitchell's novel. First, if one is interested in the identification and distinction of the chapters of Mitchell's novel, then, in fact, 92% of the distribution of words between the different sections of the text is irrelevant. This is not to say that they are not also different, just that they are more closely correlated than the 8% that act as strongly discriminative markers of each section. Second, while a conventional reader might argue that it is the unique the-

Cloud Atlas vs Moby Dick
Bootstrap Consensus Tree

David Mitchell Cloud Atlas Ghastly 1
David Mitchell Cloud Atlas Ghastly 2
David Mitchell Cloud Atlas Letters 2
David Mitchell Cloud Atlas Letters 1
Melville Bartleby
Melville Moby Dick
David Mitchell Cloud Atlas Orison 2
David Mitchell Cloud Atlas Orison 1
David Mitchell Cloud Atlas Luisa 1
David Mitchell Cloud Atlas Luisa 2
David Mitchell Cloud Atlas Ewing 1
David Mitchell Cloud Atlas Ewing 2

25-100 MFW Culled @ 0%
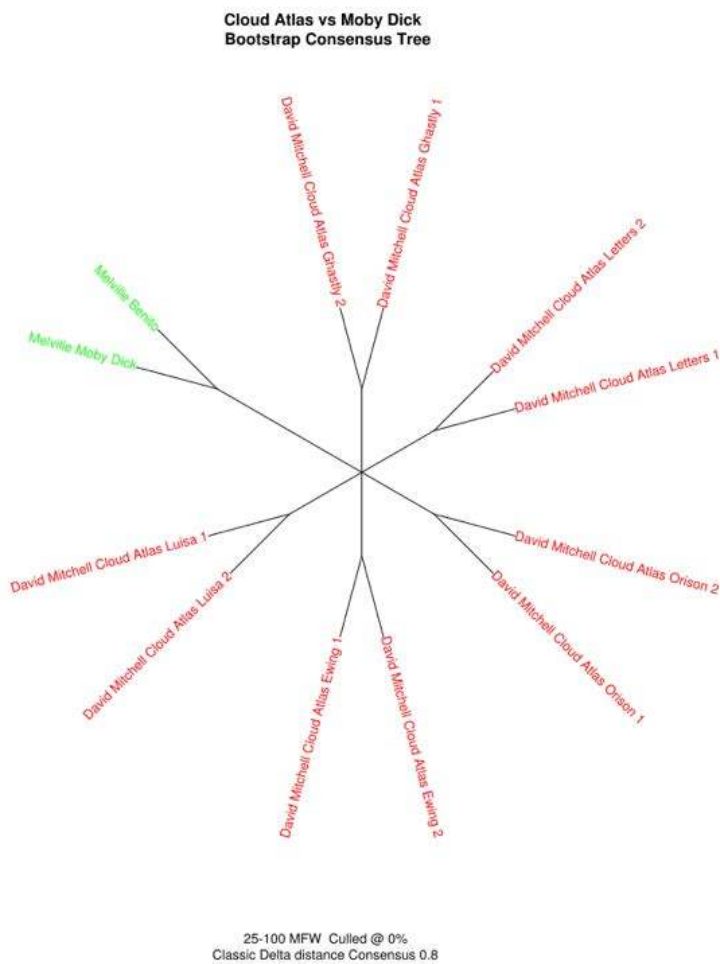Classic Delta distance Consensus 0.8

Figure 6: Melville and Mitchell compared by delta cluster bootstrap consensus tree at 0.8 consensus with 20-100 MFW.

matic and stylistic elements of each sub-text that are important ('orisons,' nuclear reactors, sea storms, retirement homes), the shifts in grammatical register that Mitchell deploys to discern his chapters from one another force perceptible micro-changes among words that usually go unobserved.

The other experiment that is worthwhile in the realm of authorship attribution techniques is to validate a character in the novel's claim that "Ewing puts me in mind of Melville's bumbler Cpt. Delano in 'Benito Cereno'" (Mitchell 1007). For, while the character may be put in mind of that text, conventional authorship attribution methods using Burrows's delta cluster Ewing with neither Melville's *Moby-Dick* nor with "Benito Cereno," using the Project Gutenburg editions (see Figure 6).

The cluster diagram in Figure 6 is particularly interesting for, while it does not show a grouping of Melville and the Ewing portion of the text with any consensus, the clustering also believes that each section of *Cloud Atlas* should be grouped independently. 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 diagram 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.' Further work that I am conducting consists of collecting various texts within the Ewing and Luisa Rey genres and profiling these against these sections to determine whether any other authors might be more closely clustered by this distance measure. In relative terms, the addition of extra texts into the clustering algorithm may also narrow the distance between the sections of Mitchell's novels, eventually resulting in an underlying authorship cluster. For now, though, Mitchell's genres are too distinct from one another, within the corpus with which I am working, to be computationally clustered.

Micro-Tectonics

These micro-tectonic, sub-surface shifts of linguistics that constitute changes to genre and register between the chapters of *Cloud Atlas* could also reasonably be expected to re-manifest in part-of-speech (PoS) trigrams. A 'trigram' refers to a set of three consecutive entities, while by 'part-of-speech' here I mean a named word type ('noun subject,' 'verb,' 'noun object,' for example, is a part-of-speech trigram). After all, the reconfiguration of the frequency of basic blocks of speech, such as deter-

miners (articles), seems likely to affect the grammatical composition of each one of the texts.

In order to investigate what might happen to Mitchell's prose within the linguistic variations of his chapters, I used the feature-rich part-of-speech tagging software known as the 'Stanford Tagger,' which uses a cyclic dependency network to assign a set of symbols to each part of speech (Toutanova et al.). Tagging parts of speech is not, however, an easy computational problem. Many words have multiple functions and are highly context dependent. This method of PoS tagging uses a set of trained models (on a broader English corpus) to look for similarities in linguistic structure and demonstrates a 97% accuracy in test runs, although I have here ignored "Sloosha's Crossin'" in my determination of accuracy. It is not likely that the tagger would work well against Mitchell's mutilated fictional language of that central chapter. The 97% accuracy benchmark, remember, means that for every 100 words of the novel, three will be misclassified.

As an example of how this tagger works, let us take the sentence "we make sail with the morning tide," which comes from the first chapter of Ewing's narrative. The Stanford tagger transforms this sentence into a symbolic dictionary of parts of speech. In this case, the output reads: 'PRP VBP VB IN DT NN NN.' Translated back into English, this means: "we [personal pronoun] make [verb, non-3rd person singular present] sail [verb, base form] with [preposition or subordinating conjunction] the [determiner] morning [noun, singular or mass] tide [noun, singular or mass]." Note here that we can see an erroneous transformation: "morning" is here actually an adjective, but is misclassified as a noun. Using the Stanford tagger, I converted each chapter of Cloud Atlas into its corresponding PoS version, yielding largely unreadable text files of the underlying linguistic structure of the novel, as determined by a 97%-accurate machine-reading apprch (Table 1).

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural

NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Table 1: a lookup table of the parts of speech produced by the Stanford tagger, here derived from the Penn classification

The first aspect that I wanted to know was whether or not PoS tagging provided another way by which we might group the chapters of *Cloud Atlas* as distinct from one another. In order to achieve this, I began by running bootstrap consensus tree imaging of the top 1,000 PoS components that occur throughout the novel, insisting that 90% of them agreed with one another in how the texts were clustered. Indeed, it does appear that in 900 of the 1000 iterations on which I performed the cluster analysis, it is possible to group the texts by the part-of-speech trigrams (Figure 7).

That said, the sensitivity of differentiation between the chapters is here far less than when using word frequency. Indeed, we cannot use the twenty most common parts of speech, for example, because there is too much overlap. In fact, there is also an insufficiently strong signal if we use only the part-of-speech trigrams that are shared between the sections of the novel. Where the text becomes interesting is when we see standout deviations of linguistic patterns that occur in certain of Mitchell's chapters and not in others.

Consider Figure 8, for example. This shows the 1,000 most-to-least common PoS trigrams throughout the text, sorted by an average across each portion of the text. It also, though, provides a useful visual index of where the texts vary from one another in terms of their unique linguistic features. If one looks approximately 1/15th of the way into the graph,
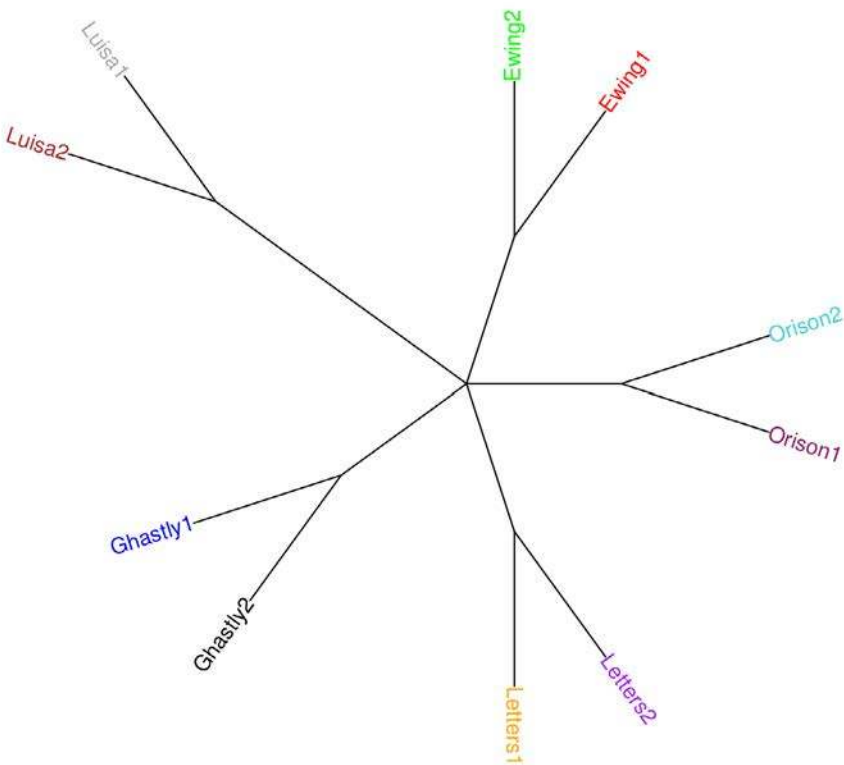
Figure 7: bootstrap consensus tree of part-of-speech tagged version of *Cloud Atlas* including all unique PoS constructs of 1,000 most common PoS trigrams.

there is one isolated point that juts out well above the others in height. This marker turns out to represent the fact that the Luisa Rey portion of *Cloud Atlas* uses the figuration NNP NNP VBZ (proper noun singular–> proper noun singular –> verb, 3rd person singular present) to a far higher extent than any of the other chapters (Table 2).

This NNP NNP VBZ formula comes about because of the Luisa Rey section's unique tendency to reuse the full name of its characters before any present-tense verb. To take but the first few instances, we can clearly see "Rufus Sixsmith leans," "Luisa Rey hears," "Maharaj Aja says," "Javier Gomez leafs," "Nancy O'Hagan has," "Jerry Nussbaum wipes," "Dom Grelsch breaks," "Joe Napier watches," "Alberto Grimaldi scans," "Isaac Sachs closes," "Roland Jakes drips," and "Bill Smoke watches," among

many other instances. While this trigram is present at around the 0.1% mark in all other chapters of *Cloud Atlas*, the Luisa Rey portion is distinct in having almost ten times as many occurrences.
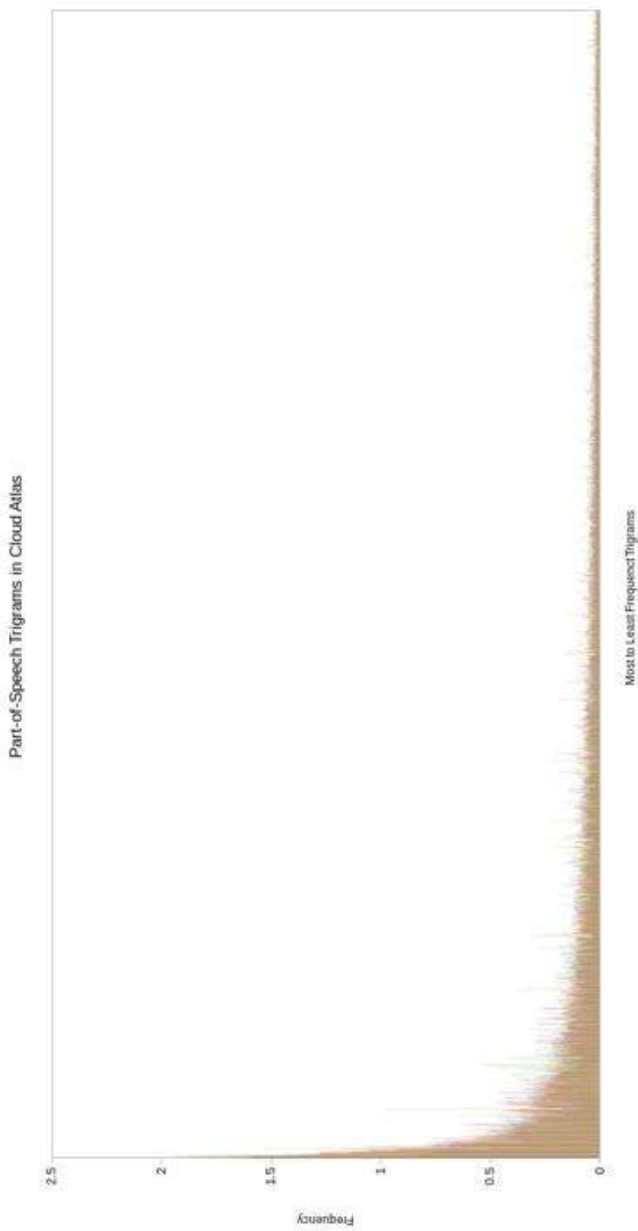

Figure 8: the 1,000 most-common PoS trigrams in *Cloud Atlas* across all sections.

PoS	Ewing	Ghastly	Letters	Luisa	Orison
nnp nnp vbz	0.16	0.16	0.12	0.97	0.06

Table 2: one of the anomalous trigrams (NNP, NNP, VBZ) in the PoS tagging of *Cloud Atlas*

While the above graph is helpful in determining which linguistic features are of interest and are unique to each section, a better way to achieve this is to calculate the standard deviation from the average frequency and to note outlier points by comparing to this. For instance, in the example I was just using, the average frequency of occurrence of the NNP NNP VBZ is 0.30. The standard deviation (that is, the average amount by which every chapter frequency for NNP NNP VBZ varies from this average) of this line is 0.33. The Luisa Rey chapter, then, at 0.97 is 1.98 standard deviations above the mean, which, assuming a normal distribution of PoS trigrams across the whole text, is in the top 5% of anomalous results. If, then, we plot the standard deviations and remove all entries from the table where no single text reaches a 1.9 standard deviation, we can plot a stacked percentage chart (Figure 9) that can serve as a strong visual index of unique part-of-speech formulations.

In this chart, the vertical width of each striated band represents the relative use of the 123 trigrams that score at a standard deviation of 1.9 as though the sum of each column were 100%. This allows us to visualize the difference between sections for each trigram without the actual frequency values between each trigram masking internal differences. In other words, columns cannot be compared to each other on an absolute basis. The fact that one column is taller than another does *not* mean that the trigrams on the right that are wider than those on the left actually occur more frequently. What it does mean is that, in relative terms, the taller the bar, the more frequently a section uses a trigram *compared to the other sections within its column*. Indeed, the results towards the right of the above graph are often the difference of only a single greater occurrence of a trigram between sections (and given that we have a 3% error rate, we should be wary here). In this sense, such results are both more *and* less reliable. They are more reliable as markers of distinction, since they occur precisely a single unit more or less than counterpart chapters, error rate notwithstanding. They are less reliable because the variance is far more

likely to have been introduced by utter chance rather than any aesthetic/ stylistic control on Mitchell's part.

Indeed, on this type of calculation and visualization, the Luisa Rey portion of the narrative presents itself as the most different to all others with 74 out of 1,000 trigrams occurring at the 1.9 standard deviation mark. For example, another formulation that is uncommon among the other parts of the novel except for Luisa Rey is VBZ DT NN (verb, 3rd-person singular present –> determiner –> noun, singular or mass). This is partly
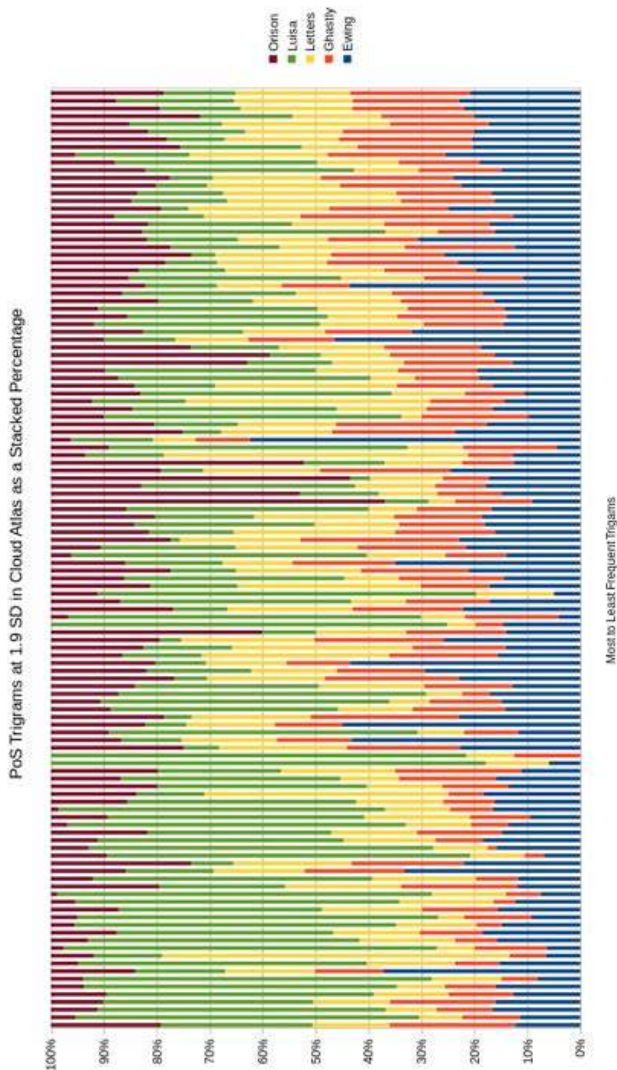

Figure 9: PoS trigrams at 1.9 standard deviations in *Cloud Atlas* as a stacked percentage chart

a result of the novel's present-tense setting and consists of formulations such as "hits the sidewalk," "slams the balcony," "hears a clunk," "shows the world," and so on. Indeed, the present tense narration of the Luisa Rey chapter gives it a unique flavor and there are many instances of VBZ-type formulations that do not exist elsewhere in the novel. For instance, we also see NNP VBZ DT (proper noun, singular –> verb 3rd person singular present –> determiner) with a much greater frequency in this chapter than elsewhere ("Luisa inspects the," "Luisa manages a," "Javier attaches the," etc.). In fact, as a general rule, the Luisa Rey segment can be said to be characteristically different from the other sections of *Cloud Atlas* in its use of present-tense narration that includes VBZ formulations occurring with 1.9 standard deviations more frequency than the average of other portions of the text. As one would expect as a correlative, many VBD (verb, past tense) formulations occur at significantly lower levels in the Luisa Rey narrative. This is clearly part of the generic distinction of the thriller formation of this portion of the novel. It is lent a fast pace by the present-tense trot of the text. The reuse of full names at the start of each chapter serves to seemingly relocate the action in a slamming fashion, a total and distinct re-placement of the reader through full-name appellation.

The next most linguistically distinct portion of *Cloud Atlas* is "The Pacific Diary of Adam Ewing," which contains fifteen trigrams that occur at over or under 1.9 standard deviations from the mean (albeit not all of which seem to distinguish the chapter from others in a reliable fashion; see above). Indeed, Ewing's narrative can be categorized as over-using IN DT NNS (preposition or subordinating conjunction –> determiner –> noun, plural), represented in formulations such as "on the stairs," "than the digits," "through the paths," "inside the coils"; DT NNS IN (determiner –> noun, plural –> preposition or subordinating conjunction), seen in "the fangs of," "the pearls of," "the works of"; NNP CC PRP (proper noun, singular –> coordinating conjunction –> personal pronoun), mostly instances of "Henry & I." Put otherwise, the Ewing narrative is linguistically distinct in order to achieve two features of its generic register and thematic concerns that are important for the text. The first is that, in the use of DT NNS IN and NNP CC PRP, the Pacific Diary narrative gives many more comparative and locative descriptions of characters and artifacts than do other portions of the text. This lends a degree of formal pedantry to the voice here that is not present elsewhere. In the second case, the NNP CC PRP formulation is integral to establish the supposed friendship with Henry Goose that leads to Ewing's near-downfall. However, the tight usage of "Henry & I" here, consistently with no slippage, contributes to the historical imaginary of the 1850s writing style as an era where grammar was 'correct' and people wrote in a formal register.

By contrast, Ewing's narrative is short on JJ JJ NN (adjective –> adjective –> noun, singular or mass) and RB JJ NN (adverb –> adjective –> noun, singular or mass). While, then, Luisa Rey's narrative contains a "hopelessly uneven gunfight," a "mostly empty wine" glass, and "very little traffic," such formulations are rare or even non-existent in the Ewing section. This lends a specificity or qualifying nuance to the Luisa Rey narrative. It is also, though, clearly a trope of hackneyed over-written airport thrillers to modify every term that is used in this way. These linguistic tropes – just some of the many that the amplifying visualization technique allows us to see – are the substrate upon which Mitchell's genre effects are built.

Seeing the Ocean for the Drops

I have attempted, in this article, to provide a demonstration of the ways in which computational methodologies can be used to garner new empirical evidence that can then be fed back into traditional close-reading and theoretical approaches. This article forms part of a longer work in progress that more extensively interprets the results from the computational microscopic/quantitative formalistic techniques that I am using. There are many more techniques to be explored here, particularly in the realm of neural networks for authorship attribution, which is a fast-growing field. What I have tried to show, though, is that digital methodologies need not be utilitarian in the ways that they approach literature. We can use these approaches in symbiosis with more conventional literary interpretation. Indeed, above, I gave some significant thought to what we mean by 'literary style,' through a questioning of the conditions under which, I contend, we frequently assume that writers work. This theorizing was made possible through the digital approaches of stylometry. I then moved to examine how we might use a computational approach to pull out significantly more common part-of-speech patterns between portions of a novel. This, in turn, opened the possibility of a more-informed linguistic criticism of Mitchell's genre techniques.

The benefits of such an approach are, then, reciprocal. Literary theory, I contend, can find itself enriched through a new set of methodologies and the cracks in our thinking that they expose. Literary criticism, on the other hand, is armed with a fresh set of observations that are difficult to spot by eye, but that can be extracted using computational techniques. In many ways, the methods I use here and that I have described as a microscope can also be understood through a different imperfect metaphor, though: filtration. As the ocean of the text is sifted for minerals that we might use, its drop-like composition at the linguistic level that causes the macro oceanic effects can be better discerned. Such a forced metaphor is, of course,

apt, for thinking about *Cloud Atlas*. For as Mitchell's Ewing closes the novel, he asks of the reader: what is any ocean but a multitude of drops?

Birkbeck, University of London

Works Cited

Allington, Daniel, Sarah Brouillette, and David Golumbia. "Neoliberal Tools (and Archives): A Political History of Digital Humanities." *Los Angeles Review of Books*, 11 May 2016. Web. 28 Aug. 2017.

Allison, Sarah, et al. "Quantitative Formalism: An Experiment." Stanford Literary Lab, 2011. Web. 17 Aug. 2016.

Argamon, S. "Interpreting Burrows's Delta: Geometric and Probabilistic Foundations." *Literary and Linguistic Computing*, vol. 23, no. 2, 2007, pp. 131-147. Web.

Barthes, Roland. "The Death of the Author." *Image, Music, Text*, translated by Stephen Heath, Fontana Press, 1987, pp. 142-148.

Brennan, Michael Robert, and Rachel Greenstadt. "Practical Attacks Against Authorship Recognition Techniques." *IAAI*, 2009. Web. 1 Aug. 2016.

Burke, Sean. *The Death and Return of the Author: Criticism and Subjectivity in Barthes, Foucault and Derrida*. 3rd Revised edition, Edinburgh University Press, 2008.

Burrows, John. "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic* Computing, vol.17, no. 3, 2002, pp. 267-287. Web.

Canter, David. "An Evaluation of 'Cusum' Stylistic Analysis of Confessions." *Expert Evidence*, vol. 1, no. 2, 1992, pp. 93-99.

Chaski, C. E. "The Keyboard Dilemma and Forensic Authorship Attribution." *Advances in Digital Forensics*, vol. 3, 2007.

---. "Who's at the Keyboard: Authorship Attribution in Digital Evidence Investigations." *International Journal of Digital Evidence*, vol. 4, no. 1, 2005.

Dimovitz, Scott. "The Sound of Silence: Eschatology and the Limits of the Word in David Mitchell's *Cloud Atlas*." *SubStance*, vol. 44, no. 1, 2015, pp. 71-91. Web.

Eder, Maciej. "Visualization in Stylometry: Cluster Analysis Using Networks." *Digital Scholarship in the Humanities*, 2015, pp. 1-15. Web.

Eder, Maciej, Mike Kestemont, and Jan Rybicki. "Stylometry with R: A Suite of Tools." *Digital Humanities 2013: Conference Abstracts*, University of Nebraska–Lincoln, 2013, pp. 487-89. Web.

"Editors' Choice: Round-up of Responses to 'The LA Neoliberal Tools (and Archives).'" *Digital Humanities Now*. Web. 11 May 2016.

Elliot, W., and R. J. Valenza. "And Then There Were None: Winnowing the Shakespeare Claimants." *Computers and the Humanities*, vol. 30, 1996, pp. 191-245.

---. "The Professor Doth Protest Too Much, Methinks." *Computers and the Humanities*, vol. 32, 1998, pp. 425-90.

---. "So Many Hardballs so Few over the Plate." *Computers and the Humanities*, vol. 36, 2002, pp. 455-60.

Eve, Martin Paul. "The Great Automatic Grammatizator: Writing, Labour, Computers." *Critical Quarterly*, 2017. Web. 28 June 2017.

---. "Scarcity and Abundance." *The Bloomsbury Handbook of Electronic Literature*. Bloomsbury Academic, 2017.

---. "'You Have to Keep Track of Your Changes': The Version Variants and Publishing History of David Mitchell's *Cloud Atlas*." *Open Library of Humanities*, vol. 2, no. 2, 2016, pp. 1-34. Web.

Evert, Stefan et al. "Outliers or Key Profiles? Understanding Distance Measures for Authorship Attribution." *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków. Web. 15 Aug. 2016.

Farringdon, J. M. *Analyzing for Authorship: A Guide to the Cusum Technique*. University of Wales Press, 1996.

Foucault, Michel. "What Is an Author?" *The Essential Works of Michel Foucault, 1954-1984*. Vol. 2. Penguin, 2000, pp. 205–222.

Gitelman, Lisa, editor. *"Raw Data" Is an Oxymoron*. The MIT Press, Infrastructures Series, 2013.

Gladwin, Alexander A. G., Matthew J. Lavin, and Daniel M. Look. "Stylometry and Collaborative Authorship: Eddy, Lovecraft, and 'The Loved Dead.'" *Digital Scholarship in the Humanities*, vol. 32, no. 1, 2017, pp. 123-140. Web.

Grieve, J. W. "Quantitative Authorship Attribution: A History and an Evaluation of Techniques." Masters, Simon Fraser University, 2005. Web.

Hardcastle, R.A. "CUSUM: A Credible Method for the Determination of Authorship?" *Science & Justice*, vol. 37, no. 2, 1997, pp. 129-138. Web.

---. "Forensic Linguistics: An Assessment of the CUSUM Method for the Determination of Authorship." *Journal of the Forensic Science Society*, vol. 33, no. 2, 1993, pp. 95-106. Web.

Hilton, M. L. "An Assessment of Cumulative Sum Charts for Authorship Attribution." *Literary and Linguistic Computing*, vol. 8, no. 2, 1993, pp. 73-80. Web.

Holmes, David I. "The Evolution of Stylometry in Humanities Scholarship." *Literary and Linguistic Computing*, vol. 13, no. 3, 1998, pp. 111–117.

Holmes, David I., and Fiona Tweedie. "Forensic Stylometry: A Review of the Cusum Controversy." *La Revue Informatique et Statistique dans les Sciences Humaines*, vol. 31, nos. 1–4, 1995, pp. 19–47.

Hoover, D. L. "Testing Burrows's Delta." *Literary and Linguistic Computing*, vol. 19, no. 4, 2004, pp. 453-475.

Hopf, Courtney. "The Stories We Tell: Discursive Identity Through Narrative Form in *Cloud Atlas*." *David Mitchell: Critical Essays*, edited by Sarah Dillon, Gylphi, 2011, pp. 105–126.

Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, 2013.

Juola, Patrick. "Authorship Attribution." *Foundations and Trends® in Information Retrieval*, vol. 1, no. 3, 2007, pp. 233-334. Web.

---. "Stylometry and Immigration: A Case Study." *JL & Pol'y*, vol. 21, 2012, p. 287.

Kenny, Anthony. *The Computation of Style: An Introduction to Statistics for Students of Literature and Humanities*. 1st ed., Pergamon International Library of Science, Technology, Engineering, & Social Studies, Pergamon Press, 1982.

Kerr, Norbert L. "HARKing: Hypothesizing After the Results Are Known." *Personality and Social Psychology Review*, vol. 2, no. 3, 1998, pp. 196-217. Web.

McMenamin, G. "Disputed Authorship in US Law." *International Journal of Speech, Language and the Law*, vol. 11, no. 1, 2004, pp. 73-82.

Mitchell, David. *Cloud Atlas*. Kindle edition, Random House, 2008.

Moretti, Franco. *Distant Reading*. Verso, 2013.

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

---. "The Slaughterhouse of Literature." *MLQ: Modern Language Quarterly*, vol. 61, no. 1, 2000, pp. 207-227.

Morton, A. Q. *Literary Detection: How to Prove Authorship and Fraud in Literature and Documents*. Bowker, 1978.

Mosteller, F., and D. L. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, 1964.

O'Donnell, Patrick. *A Temporary Future: The Fiction of David Mitchell*. Bloomsbury Academic, 2015.

Pearl, Lisa, Kristine Lu, and Anousheh Haghighi. "The Character in the Letter: Epistolary Attribution in Samuel Richardson's *Clarissa*." *Digital Scholarship in the Humanities*, vol. 32, no. 1, 2016, pp. 123-140. Web.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2016. Web.

Rybicki, J., and M. Eder. "Deeper Delta across Genres and Languages: Do We Really Need the Most Frequent Words?" *Literary and Linguistic Computing*, vol. 26, no. 3, 2011, pp. 315-321.

Saïd, Edward W. *On Late Style*. Bloomsbury, 2006.

Shanahan, John. "Digital Transcendentalism in David Mitchell's Cloud Atlas." *Criticism*, vol. 58, no. 1, 2016, p. 115. Web.

Shoop, Casey, and Dermot Ryan. "'Gravid with the Ancient Future': Cloud Atlas and the Politics of Big History." *SubStance*, vol. 44, no. 1, 2015, pp. 92-106. Web.

Stolerman, Ariel et al. "Breaking the Closed-World Assumption in Stylometric Authorship Attribution." *Advances in Digital Forensics X*, vol. 433, edited by Gilbert Peterson and Sujeet Shenoi, Springer, 2014, pp.184-205. Web. 31 July 2016.

Toutanova, Kristina et al. "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network." *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, Association for Computational Linguistics, 2003, pp. 173-180. Web. 13 June 2016.

Underwood, Ted. "The Life Cycles of Genres." *Journal of Cultural Analytics*, 2016. Web. 28 June 2017.