‘Mind our mouths and beware our talk’: Stylometric analysis of character dialogue in *The Darjeeling Limited*

Warren Buckland, Oxford Brookes University

Abstract

Stylometry uses statistical reasoning to quantify the linguistic attributes of written texts. In this article I draw upon current developments in computer-based stylometric studies to quantify the language of screenplays. I take as my starting point J. F. Burrows’s seminal stylometric study of dialogue in Jane Austen’s novels (*Computation into Criticism* [Burrows 1987]) to identify and quantify the linguistic habits of major screenplay characters, habits that constitute their distinctive voice. Analysis of the dialogue of the three Whitman brothers in *The Darjeeling Limited* (screenplay by Wes Anderson, Roman Coppola and Jason Schwartzman, dated 22 November 2006) will serve as a preliminary case study. I aim to use the work of Burrows as the starting point in establishing a new research programme within screenplay studies, one based on the stylometric analysis of the language of screenplays.

Keywords

dialogue

stylometry

John Burrows

Computation into Criticism

Wes Anderson

The Darjeeling Limited

In his seminal manual on film directing, Michael Rabiger encourages trainee directors to watch films carefully or, as he put it, to read *from* the film rather than to read *into* it. This is because

> [film] can make you uneasy about your perceptions and too ready to accept what should be seen or should be felt. Recognize what the film made you feel, then trace your impression to what can actually be seen and heard in the film. (Rabiger 2003: 80)

This simple piece of advice – 'trace your impression to what can actually be seen and heard in the film' – is not sufficiently heeded by many film students and film scholars alike, who either write impressionistic criticism or rush into an interpretation of a film. We first need to pause and focus on what is in the film.

We can apply this simple piece of advice to the analysis of screenplays, focusing on what is actually in the screenplay, the words that compose its textuality. The most extreme and exacting form of this attention to the textuality of a text is the discipline of stylometry (which overlaps with quantitative linguistics, corpus linguistics and digital literary studies). A simple definition of stylometry is a discipline that uses statistics to quantify style, a type of study that has become more feasible in the last 30 years with computing and simple software such as Excel, Voyant Tools, Textalyser and Coh-Matrix, among others. (On Voyant Tools, see Rockwell and Sinclair [2016], and for more on Coh-Metrix, see McNamara et. al. [2014].)

In this article I examine the viability of a stylometric analysis of screenplays, guided by one of the most famous studies in recent years, J. F. Burrows's book *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method* (1987). More specifically, Burrows studies the dialogue in Jane Austen's six completed novels, examining them using a variety of statistical tests (outlined below). The basic premise of this article is that Burrows's meticulous statistical analysis of dialogue in Austen's novels can provide exact methods to generate insights into the dialogue of screenplays. I present an overview of Burrows's methods and results, and test his methods using the written dialogue of the three brothers in *The Darjeeling Limited* (screenplay by Wes Anderson, Roman Coppola and Jason Schwartzman, dated 22 November 2006). This article is exploratory and primarily illustrative to the extent that my aim is to try out on one Wes Anderson film the statistical methods Jane Austen's dialogue was subjected to, in order to determine the viability of Burrows's methods for analysing screenplays.

The analysis of character dialogue in screenplays is not new. Other studies have examined closely the dialogue of screenplays, employing textual analysis, linguistics and occasionally statistics. Jill Nelmes examines the codes of realism embedded in dialogue, codes that aim 'to draw the audience into the storyworld and to develop character' (2011: 217). In Chapter 8 of *The Screenplay: Authorship, Theory and Criticism* (2010), Steven Price reviews various contemporary linguistic theories (deixis, speech act theory, Jakobson's communication model of language, comprising six factors and six corresponding functions), and assesses their value in analysing film dialogue via a series of detailed case studies (*Pulp Fiction*, *The Usual Suspects*, *Glengarry Glen Ross*). In addition contributors to *Telecinematic Discourse:*

Approaches to the Language of Film and Television Series (Piazza et al. 2011) also employ various linguistic theories (pragmatics, multimodal analysis, stylistics, corpus linguistics) to examine the different types of language used in cinema and television, with an emphasis on different styles of dialogue (serial killer speeches in horror films, humour in TV sitcoms, emotional discourse, incomprehensible and impolite dialogue), with three authors in particular (Fabio Rossi, Rose Ann Kozinski and Michael Toolan) adopting a quantitative approach.

Within digital literary studies, Jonathan Culpeper uses statistical methods to quantify and analyse the dialogue of six characters in *Romeo and Juliet*, focusing on statistically significant grammatical and lexical character patterns (in Hoover et al. [2014]: 9–34). In the same volume, David L. Hoover takes Burrows's study of Jane Austen's dialogue as a starting point to compare character voices in Wilkie Collins's novel *The Moonstone* (Collins 1868) – where multiple narrators tell parts of the same story – and Hannah Webster Foster's epistolary novel *The Coquette* (Webster Foster 1797), and concludes that the characters in *The Moonstone* are clearly distinguished in terms of their language, whereas in *The Coquette* the characters' voices are not demarcated clearly (in Hoover et al. [2014]: 64–89). In Chapter 5 of *Style, Computers, and Early Modern Drama* (a book influenced by and dedicated to John Burrows), Hugh Craig and Brett Greatley-Hirsch (2017) analyse the 50 most common function words in the dialogue of 243 plays performed in London between 1580 and 1644 to identify any changes in language over this period. They conclude that the plays change stylistically 'from more explicit, more formally patterned dialogue to more detached commentary and more anaphoric exchanges focusing on shared material' (Craig and Greatley-Hirsch 2017: 153). This is evident, in part, in the gradual

reduction over time in the number of prepositions (and the nouns they serve) and a gradual increase in auxiliary verbs: 'Overall, dialogue with an abundance of these auxiliary verbs – and with a scarcity of prepositions – will have a focus on immediate interactions, with characters referring familiarly to themselves and to those on stage and in their immediate circle' (Craig and Greatley-Hirsch 2017: 154). The work of Culpeper, Hoover, Craig and Greatley-Hirsch is representative of a growing body of research into the quantitative study of literary and dramatic texts, research that became feasible with the publication of Burrows's *Computation into Criticism* in 1987.

Data collection

What aspects of a written text can be quantified or understood numerically? Within stylometry the following textual parameters have been quantified: sentence length, pronouns, function (grammatical) words, content (lexical) words, synonymous word pairs (on/upon, while/whilst, etc.), intensifiers/hedges, modal verbs, etc. In an analytical language such as English, function words (*on, to, in, of,* etc) are important because they signal grammatical relations (unlike synthetic languages, which signal grammatical relations via inflection). In studying common function words, it is not simply their presence or absence that defines style because they are present in the work of all authors. Instead, it is their frequency. In more technical terms, it is not a matter of possessing or not possessing attributes but a matter of identifying patterns of variability in the frequency of those attributes. This is one of the main goals of statistics: to study patterns of variability in data (which makes it pertinent to the study of style), and to reduce huge amounts of data to a manageable size. Furthermore, it is not simply a matter of a few attributes, but dozens of attributes. There is no fixed

universal quantitative test for style; one needs to apply numerous tests to a text to determine the text's dominant and distinctive stylistic attributes.

One of the major premises of stylometry is that style can be defined in terms of the measurement and statistical analysis of common word types. In *Computation into Criticism* Burrows focuses on the 30 most common word types (see Table 1). But these 30 word types account for 40 per cent of all the token words – that is, all the dialogue spoken by Jane Austen's characters – and they have the merit of being used by almost all characters, which makes comparison feasible: comparison between the speaking styles of characters in the same novel and between a character and the novel's statistical average, plus the standards of language, as represented in the British National Corpus database, although several authors draw attention to the problems with comparing spontaneous speech and the artificially constructed dialogue of screenplays and novels. For example: 'Film dialogue is a complex mix of the everyday and the poetic; it creates the illusion of being natural by using colloquial words yet its actual language construction is anything but; each word is carefully chosen, more artificial than natural, a contrivance which aspires to seem real but is not' (Nelmes 2011: 236). In addition several chapters in Piazza, Bednarek and Rossi (2011) also address the relation between dialogue and spontaneous speech.

a	her	that (conj)
and	in	the
as	I	to (inf)
at	is	to (prep)
be	it	very
but	me	was
do	my	will (vb)
for (prep)	not	with
have	of	you
he	she	your

Table 1: John Burrows's list of the 30 most common word types in Jane Austen's dialogue (arranged alphabetically).

Table 1: John Burrows's list of the 30 most common word types in Jane Austen's dialogue (arranged alphabetically).

Burrows defends the stylometric analysis of literary texts by arguing that:

> However narrow the linguistic function of words like these, it is evident that if, as is indeed the case, disparities like these are typical of the language of Jane Austen's major characters, the effects must colour every speech they make and leave *some* impression in the minds of her readers. […] Statistical analysis of the peculiarities of incidence makes it possible to approach the whole penumbra of 'meaning' in a new and fruitful way. (Burrows 1987: 4 original emphasis)

Even though they are inconspicuous, these function words constitute the fabric of the text, its constituent parts. Authors tend to select function words intuitively rather than consciously; these words therefore constitute a high-frequency invariant habit rather than a conscious choice influenced by context or subject matter. We can see from Burrows's definition that, within stylometry, the term 'style' is not conflated with deliberate choice. We find a similar emphasis in the connoisseurship of Giovanni Morelli and Bernard Berenson, who attributed authorship to paintings based on small, inconspicuous and nonconscious details such as the way earlobes or finger nails were painted (see Wollheim 1973; Ginzburg and Davin 1980).

Burrows brings to the analysis of dialogue the practice of comprehensive data collection and meticulous analysis of that data based on well-established statistical techniques. He uses computing and statistics to examine quantitatively the vocabulary of character dialogue in Jane Austen's novels, especially in terms of their idiolect. From a statistical perspective, idiolect does not refer to the irreducibly singular characteristic of a language speaker; instead, it refers to a speaker's pattern of deviance from the norms of a language that is social in nature. Burrows's data consist of the dialogue of all the main characters in Austen's six novels (which he defines as those who speak 2000 words or more, adding up to 48 characters in total). He also creates a control group consisting of novels by Henry James (*The Awkward Age*), E. M. Forster (*Howard's End*), Georgette Heyer (*Frederica*) and Virginia Woolf (*The Waves*), plus Austen's manuscript fragment 'Sanditon' and the extension of the fragment into the novel *Sanditon*. In total, he analyses twelve novels, consisting of one and a quarter million words.

Through a series of empirical tests Burrows establishes that Jane Austen's main characters are defined in terms of their speech: 'the idiolects of many of Jane Austen's major characters are firmly and appropriately differentiated' (1987: 69). Burrows interprets this ability to differentiate characters using just a few thousand words to be a mark of Jane Austen's immense literary talent.

One potentially challenging issue that the work of Burrows raises for scholars in the humanities is that the literary and linguistic properties of a text are translated into numerical properties, which are then subjected to statistical reasoning. But in the era of big data, where massive data sets are collected, stored and accessed electronically,

statistical reasoning offers one way to get to grips with data by reducing and quantifying it, and by discovering within it patterns of variability or distinct groupings that humans are unable to perceive. In combining computing and statistics, stylometry can therefore generate insights about language that exceed the memory and critical analysis of any reader for it can identify unseen patterns in huge quantities of data and can make comparisons more precise. Burrows's book presents a successful and rewarding way into the statistical analysis of written texts (especially dialogue) and can assist in expanding screenplay studies to include a specific type of knowledge that only statistics can generate.

Overview of *Computation into Criticism* and its applicability to *The Darjeeling Limited*

Burrows discusses the following statistical tests in *Computation into Criticism*: chi-squared test, the mean, standard deviation, normal distribution curves, z-scores, frequency distribution (word frequency counts), correlation coefficient, linear regression, correlation matrix, principal component analysis (eigenvalues), time-series analysis, and coefficient of variation (standard deviation/mean). In this article I focus only on frequency counts, regression and the correlation coefficient, illustrating each test with data from Burrows's analysis of Jane Austen's dialogue, followed by my own data from the dialogue of *The Darjeeling Limited*. The aim is to answer the following question: what data and insights can a stylometric analysis of screenplay dialogue generate?

(1) Frequency counts

To ease humanities scholars into his book, Burrows presents a very small sample of his word counts from *Northanger Abbey*. The count of frequently used common words sets the expectation that their distribution will be evenly shared amongst the characters:

> he who speaks a fifth of all the words might be expected to employ about a fifth of all the instances of inert words like 'of' and 'the.' And yet, when each character's actual share of such words is compared with his share of the whole dialogue, there is often a gulf between the expectation and the fact. (Burrows 1987: 3)

As style is defined in terms of frequency distributions of a text's linguistic features – or patterns of variation in the frequency distributions – the gap between expectation (average distribution) and actual (observed) frequency becomes a stylistic feature of the text. In *Northanger Abbey*, Burrows informs us, the variation in the distribution of four words ('the', 'of', 'I', 'not') between Catherine Morland and Henry Tilney is stylistically significant (see Table 2).

	Catherine	Henry
the	16.34	35.29
of	15.91	29.92
I	56.68	24.56
not	26.99	12.69

Table 2: The variation in the distribution of four words (the, of, I, not) between Catherine Morland and Henry Tilney in Northanger Abbey *(words per 1000).*

Table 2: The variation in the distribution of four words (*the, of, I, not*) between Catherine Morland and Henry Tilney in *Northanger Abbey* (words per 1000).

To make comparison possible, Burrows standardizes the data by representing each word frequency in terms of rates per 1000 (which can be divided by ten to represent them as percentages). Catherine's frequencies for 'the' (16.34 per 1000) and 'of' (15.91 per 1000) are half those of Henry (35.29 and 29.92 per 1000, respectively), while her uses of 'I' (56.68) and 'not' (26.99) are more than double his frequencies (24.56 and 12.69, respectively). From this type of data he argues that '*[f]rom no other evidence* than a statistical analysis of the relative frequencies of the very common words, it is possible to differentiate sharply and appropriately among the idiolects of Jane Austen's characters' (1987: 4, original emphasis). Burrows also discusses word counts of pronoun use in his entire sample in Chapter 1, and spends several pages examining the use of first-person plural pronouns in the dialogue of Jane Austen's characters.

What insights can frequency counts bring to our understanding of screenplay dialogue? *The Darjeeling Limited* is centred on the three Whitman brothers: Francis (Owen Wilson), Peter (Adrien Brody) and Jack (Jason Schwartzman). They are traumatized by the death of their father and the disappearance of their mother. Accompanied by a soundtrack that includes three songs from The Kinks, from where the quotation in the title of this article derives (*Strangers*, track two of *Lola Versus Powerman and the Moneygoround, Part One* [1970]), Francis organizes a train journey around India to bring his brothers together and to locate their mother.

From the screenplay I created three separate files ('Francis', 'Peter', 'Jack'), each one containing the dialogue of each brother (manually extracted via cutting and pasting). I then uploaded each file to 'textalyser.net', which generates several statistical tables,

including an overview of each file and a frequency count of words in each file. Table 3 combines the overview of the three files and Table 4 combines a small sample of the word frequencies of each file.

	Francis	Peter	Jack
Total word count:	2561	1142	955
Number of different words:	713	412	359
Complexity factor (Lexical Density):	27.8%	36.1%	37.6%
Readability (Gunning-Fog Index): *(6-easy 20-hard)*	3.3	2.5	2.2
Average Syllables per Word:	1.41	1.41	1.38
Sentence count:	406	226	211
Average sentence length (words):	6.32	5.05	4.53

Table 3: Overview of the dialogue from the three Whitman brothers in The Darjeeling Limited *(data generated from textalyser.net).*

Table 3: Overview of the dialogue from the three Whitman brothers in *The Darjeeling Limited* (data generated from textalyser.net).

Word	Francis Occurrences (Frequency)	Peter Occurrences (Frequency)	Jack Occurrences (Frequency)
and	56 (2.2%)	7 (0.6%)	2 (0.2%)
did	9 (0.4%)	1 (0.1%)	5 (0.5%)
didn't	8 (0.3%)	7 (0.6%)	3 (0.3%)
do	16 (0.6%)	9 (0.8%)	5 (0.5%)
don't	15 (0.6%)	14 (1.2%)	15 (1.6%)
have	47 (1.8%)	3 (0.3%)	4 (0.4%)

Table 4: Sample word frequencies from the dialogue of the three Whitman brothers in The Darjeeling Limited *(data generated from textalyser.net).*

Table 4: Sample word frequencies from the dialogue of the three Whitman brothers in *The Darjeeling Limited* (data generated from textalyser.net).

Table 3 reveals that Francis, who plans and leads the expedition in India, speaks more words than his two brothers combined. 'Complexity factor' measures the ratio between lexical (or content) words and function words. Francis has the lowest ratio

(that is, the lowest percentage of content words). But all the dialogue is extremely easy to read according to the Gunning-Fog Index, which designates a score of 6 as easy and 20 as hard. The brothers' scores are half that, ranging from 2.2 to 3.3. The syllable count per word is very small for all three brothers, as is the average sentence length – ranging from 4.53 to 6.32 words per sentence. These are low figures for what is considered a 'smart' and sophisticated art house film, but it demonstrates that many characters speak in fragments or just utter a few words at a time. Contrast these numbers to the mean length of sentences in Jane Austen's dialogue, which Burrows calculates to be 15.7 words (1987: 213). These simple preliminary results from the dialogue of *The Darjeeling Limited* confirm Steven Price's observation that, in comparison to the theatre (and the novel), 'the greater visual flexibility of cinema means that dialogue tends to be more compressed in the screenplay' (2010: 147).

In Table 4 I have selected from the word frequency tables a handful of words that mark Francis's dialogue as distinctive. Just as Jane Austen was able to create a distinct idiolect for Harriet Smith with a high frequency of 'and' and 'I', a low frequency of 'my' and 'to', and distinct vocabulary items 'now', 'really' and 'almost' (see Burrows 1987: 117), the screenwriters of *The Darjeeling Limited* have created a distinct idiolect for Francis. In comparison to Peter and Jack, Francis's dialogue is marked by a high frequency of 'and' and 'have'. The word 'and' is a connective; it has an accumulative function that joins clauses and sentences together, usually in a loose casual formation. Its high frequency is therefore an indication of a distinctive grammatical pattern in Francis's idiolect. Indeed, he uses 'and' to create long sentences of actions, such as the re-telling of his motorcycle accident or his list of planned activities. Speaking of Brendan's duties, Francis says to his brothers: 'He's

going to give us an updated schedule under our doors every morning of all the temples *and* spiritual places we need to see *and* expedite the hotels *and* transportation *and* everything' (emphasis added).

In contrast to his use of 'and' 56 times, Francis only uses the subordinate 'which' twice to link together clauses. (The subordinate creates a hierarchy between two clauses, whereas 'and' simply conjoins them.)

'Have' expresses multiple meanings, from functioning as an auxiliary verb that can carry tense, to forming the beginning of a question (usually associated with a pronoun), or it can signify obligation or possession. Francis uses it primarily to indicate possession ('I'm going to have the chicken', 'I'm going to have the pudding') or in questions (in combination with 'you'): 'Do you have any power adaptors?'; 'Do you have any questions?' 'Have you heard anything from Mom?'

In regard to the verb 'do', there is very little variation in terms of its positive use between the three brothers. Combining their results for 'do' and 'did' yields near identical percentages for each (Francis: 1%; Peter 0.9%; Jack: 1%). But combining their results for the two negative terms ('did not', 'do not') yields a distinct difference (Francis: 0.9%; Peter 1.8%; Jack: 1.9%). Unlike his brothers, Francis tries to sound less negative for his negative use of 'do' is half that of his brothers.

The dialogue in *The Darjeeling Limited* also contains a large number of question marks:

Whole screenplay: 303

Francis: 119 (30% of his sentences)

Peter: 50 (22% of his sentences)

Jack: 65 (31% of his sentences).

In Jane Austen, Catherine (*Northanger Abbey*) asks the highest percentage of questions (630 sentences, 138 questions: 21.9 per cent), closely followed by Lady Catherine in *Pride and Prejudice* (205 sentences, 44 questions: 21.46 per cent). Questions have multiple functions: the character can simply be seeking information (Jack's first line is simply: 'Have you seen Francis?'), whereas other questions are seeking assurances from their interlocutor. Francis is always seeking assurances from his brothers – he asks them on several occasions: 'Can we agree to that?', 'How does that sound?' and 'Do you trust me?' Francis's precarious character trait therefore emerges in part from this specific use of the question mark in his dialogue.

I mentioned above that in Chapter 1 of *Computation into Criticism* Burrows discusses pronouns in the dialogue of Jane Austen's characters, especially first-person plural pronouns. There are three first-person plural pronouns in English (or four, if we include the reflexive 'ourselves'): 'we' (nominative case – the subject of the sentence); 'us' (objective/accusative case – the direct object of a verb); and 'our' (possessive/genitive case).

In general, these pronouns allow the speaker to associate themselves with others, to include themselves in a group and its shared goals and motives. After counting the frequency of these pronouns, Burrows takes as an example Miss Bates in Jane

Austen's novel *Emma*: 'Miss Bates is the most given, in relative terms, to using pronouns in the first-person plural' (1987: 30). He then identifies the types and frequencies of first-person pronouns that she uses: 'she is much less given to the use of "our" […] than to the use of "we"; and […] her use of "us" towers above both' (Burrows 1987: 30). Miss Bates therefore prefers 'us', the first-person plural pronoun in the objective/accusative case. Burrows speculates: 'The difference appears to arise from idioms that tend to objectify the family group for which she customarily speaks and tend to acknowledge a certain passivity or submissiveness as part of its inevitable role' (1987: 31).

Table 5 shows the frequency distribution of the first-person plural pronouns in the dialogue of the three brothers in *The Darjeeling Limited* (extracted from the three separate analyses of the brothers' dialogue).

First Person Plural Pronoun	Francis Occurrences (Frequency)	Peter Occurrences (Frequency)	Jack Occurrences (Frequency)
our	5 (0.2%)	3 (0.3%)	1 (0.1%)
us	45 (1.7%)	4 (0.3%)	6 (0.6%)
we	36 (1.4%)	12 (1%)	10 (1%)

Table 5: Frequency distribution of the first-person plural pronouns in the dialogue of the three brothers in The Darjeeling Limited.

Table 5: Frequency distribution of the first-person plural pronouns in the dialogue of the three brothers in *The Darjeeling Limited*.

We can see that, generally, Francis uses the three first-person plural pronouns at a greater rate than his brothers. The differences in the rate of use of 'our' are insignificant. He uses 'we' almost twice as much as his brothers. But the 'us' score is the most significant: he uses it five and a half times more than Peter, and three times

more than Jack. (This includes all incidents of ‘Let’s’, which is a contraction of ‘let us’.)

One explanation is that Francis is driven by the desire to reunite his estranged brothers, which is reflected in his language. He takes control and organizes the trip, along with his assistant Brendan. Of the 45 times Francis says ‘us’, twenty of them are in the contracted form ‘Let’s’, all of which appear at the beginning of each sentence:

> Let’s do it.
>
> Let’s get a shoe-shine
>
> Let’s get high.
>
> Let’s get into it!
>
> Let’s go get a drink and smoke a cigarette. (*twice*)
>
> Let’s go home.
>
> Let’s go!
>
> Let’s look at the itinerary. (*three times*)
>
> Let’s make an agreement.
>
> Let’s make another agreement: […] (*three times*)
>
> Let’s see. (*twice*)
>
> Let’s set aside the next ten minutes […]
>
> Let’s update me. (*twice*)

In contrast, Jack and Peter begin sentences with ‘let’s’ just twice each.

Francis's speech is repetitive, starting sentences in the same way or using several phrases more than once. But this repetition is not the only notable feature of Francis's use of pronouns. Let's focus briefly on the final phrase on this list, where Francis is addressing his assistant Brendan: 'Let's update me' (used twice). This simple sentence combines 'us' and 'me'. In other words, Francis includes himself as both the subject and the object of the sentence.

On another occasion, the train stops, for it is lost, and passengers get off. Francis asks Brendan what is happening. Brendan says in a neutral tone: 'We haven't located us yet'. Francis pounces on this sentence and repeats it very slowly. He takes it to be symbolic, no doubt because it sounds mysterious – not only because the idea of a train getting lost is unusual but also because the sentence combines 'we' and 'us' in the same sentence; in other words, the same group is both subject and object of the sentence. (This double positioning of the speaker as both subject and object is common in reflexive sentences – e.g., 'I hurt myself' – but these sentences in *The Darjeeling Limited* are not grammatically reflexive.)

One more example of Francis's unusual use of pronouns: All three brothers are addicted to prescription medication. On one occasion Francis says to Jack: 'You're a drug addict – all of us!' Here we see Francis shifting in the same sentence from second-person singular ('you') to first-person plural ('us') to include himself with his brother; he does not want to isolate Jack. (Of course, the relation between the three brothers shifts as the story progresses; a time-series analysis – which Burrows carries out in Chapter 10 of *Computation into Criticism* – can capture the dynamics of the brothers' complex relationship in more detail.)

In sum, Francis's goals, motives and personality traits are in part expressed in the way he uses pronouns. We have seen that plural pronouns constitute a collective idiom that multiplies the participants in the utterance, enabling speakers to associate themselves with others. In *The Darjeeling Limited*, Francis's attempt to create unity amongst his brothers is signified in the frequent use of the pronoun 'us' in his idiolect, the first-person plural pronoun in the objective/accusative case. Yet, this strategy creates tension for he also ends up speaking for all three of them. Simple statistical analysis isolated Francis's use of plural pronouns in his idiolect, making comparison with his brothers' idiolects possible. In more general terms, the profile of Francis's idiolect presented over the previous pages (number of words spoken, readability, average sentence length, high frequency of 'and' and 'have', high number of questions seeking assurances, use of first-person plural pronoun 'us', repetition of phrases) confirms Burrows's claim that '[s]tatistical analysis of the peculiarities of incidence makes it possible to approach the whole penumbra of "meaning" in a new and fruitful way' (Burrows 1987: 4).

(2) Linear regression and (3) the correlation coefficient

A linear regression graph represents a relationship between two variables (two words, two characters, etc.) plotted on x and y axes. A straight line (the regression line or the line of best fit) passes through the graph, representing the ideal linear relation between the two variables. (Burrows sometimes uses a true diagonal, which serves the same function.) On a linear regression graph we can immediately see the relation between the variables plotted on the graph and the regression line. The closer the variables are to the line, the closer they are correlated.

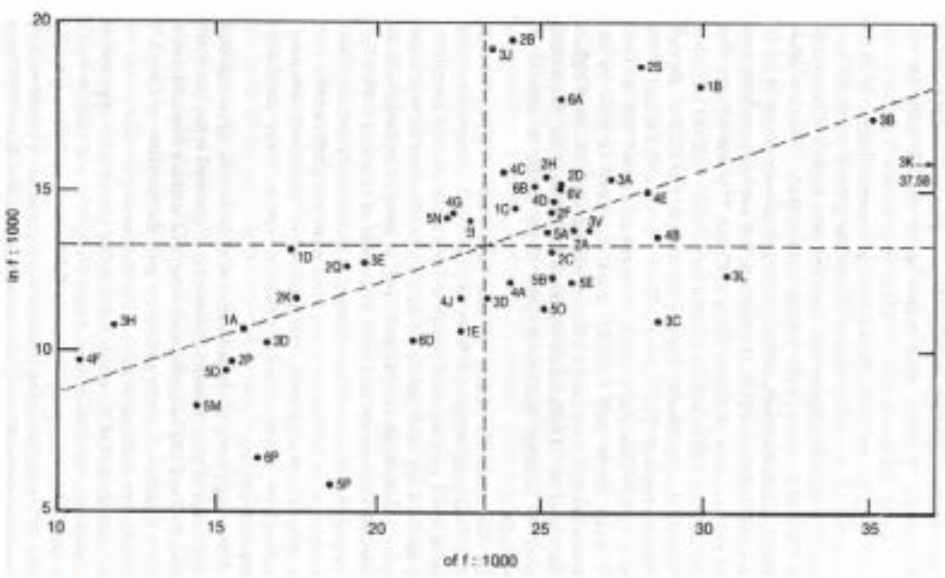
The correlation or linear association between two variables can also be expressed numerically – the correlation coefficient. Values of the correlation coefficient are always between –1 and +1. Values close to +1 signify a strong positive correlation between the two variables (if one increases the second will increase at a similar rate). Values close to –1 signify a strong negative or inverse correlation between the two variables (if one increases the second will decrease at a similar rate). If the value is 0 there is no correlation.

Burrows uses graphs and correlation coefficients to show which characters are close and which are distant from each other in terms of their vocabulary, or idiolect. One general observation that Burrows makes in regard to Jane Austen's dialogue is that the correlation coefficient between characters' usage rates of the 30 most common words is positive and typically very high (above 0.7).

In Chapter 4 of *Computation into Criticism*, Burrows studies the relationships between pairs of words, which he expresses in terms of the frequency distribution of words per 1000. He uses linear regression graphs with a line of best fit and also calculates the correlation coefficient. In one experiment, he calculates the relation between the words 'very' and 'quite' as used by all 48 characters (Burrows 1987: 63–69), and in another he focuses on the prepositions 'of' and 'in', again as used by all 48 characters (Burrows 1987: 69–75).

Burrows represents the relation between the two variables 'of' and 'in' in a linear regression graph (see Graph 1). In numerical terms, the correlation coefficient

between ‘of’ and ‘in’ in the dialogue of all 48 characters is 0.627 (Burrows 1987: 70), a fairly strong positive correlation (which means that, when the frequency of one word increases, the other tends to increase as well; the pattern between the variables is therefore positive and broadly linear). For ‘very’ and ‘quite’, the coefficient is only 0.316 (Burrows 1987: 59) – that is, weak positive correlation.



Graph 1: The relation between the two variables ‘of’ and ‘in’ in the speech of 48 Jane Austen characters (from Burrows 1987: 71). © Oxford University Press. Used with permission.

Graph 1: The relation between the two variables ‘of’ and ‘in’ in the speech of 48 Jane Austen characters (from Burrows 1987: 71). © Oxford University Press. Used with permission.

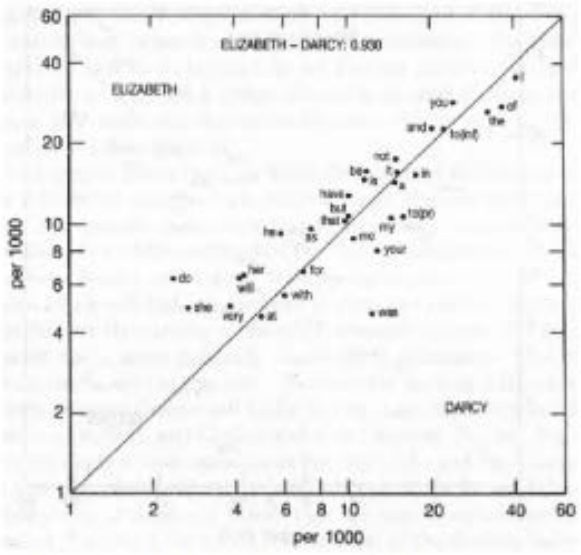
The regression line (line of best fit) passes through the graph at a diagonal. Each point on the graph represents a major character in Jane Austen’s novels (48 in total). The graph visually represents the correlation between the frequency of their use of the word ‘of’ (x-axis) in relation to their use of the word ‘in’ (y-axis). If a point (representing a character) appears on the regression line, it means that the character

uses the two words with the exact same frequency. For example, 1A (Catherine Morland) is on the regression line, which means that she uses 'of' and 'in' at the same rate. 4E (Sir Thomas Bertram from *Mansfield Park*) is also on the line, but higher up, which means that he uses both words at the same rate but more frequently than Catherine Morland. William Collins (3K), the complacent, self-conceited clergyman in *Pride and Prejudice*, is almost off the graph due to the high frequency of his use of 'of' (see top right of the graph).

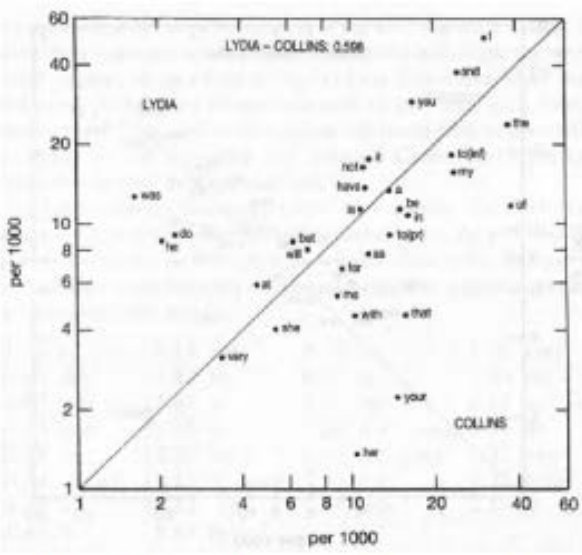
Why is this significant? Burrows argues that the variability in the use of prepositions such as 'of' and 'in' by many of Jane Austen's major characters can distinguish one character from another. This is because the study of prepositions tells us something about a character's use of grammar. In other words, the distribution of frequencies demonstrates that the study of prepositions 'is largely due to characteristic differences of syntax' (Burrows 1987: 70). Burrows presents the dialogue of William Collins as an example: 'For Collins, the "in" that governs vague abstractions figures, on average, in every second sentence that he speaks. And "of," chiefly used for the "post-modification" of his cherished abstract nouns, figures in almost every sentence' (1987: 73–75). Burrows then quotes an example of Collins's overuse of 'of': '"I am by no means *of* opinion, I assure you," said he, "that a ball *of* this kind, given by a young man *of* character, to respectable people, can have any evil tendency."' (Collins, quoted in Burrows 1987: 75, emphasis added)

Burrows interprets this as 'the absurd pomposities of Collins' (1987: 73). (We should note that, in addition to the three uses of 'of' in one sentence, Collins also uses 'by' twice.)

In Graph 1 a given pair of words in relation to the 48 characters is represented in terms of a regression line and correlation coefficients. In Chapter 5 Burrows reverses this procedure: a given pair of characters is compared in relation to their use of the 30 most common words.



Graph 2: Elizabeth and Darcy (correlation of word-types 1–30) (from Burrows 1987: 83). © Oxford University Press. Used with permission.



Graph 3: Collins and Lydia (correlation of word-types 1–30) (from Burrows 1987: 84). © Oxford University Press. Used with permission.

Graph 2: Elizabeth and Darcy (Correlation of word-types 1–30) (from Burrows 1987: 83). © Oxford University Press. Used with permission.

Graph 3: Collins and Lydia (Correlation of word-types 1–30) (from Burrows 1987: 84). © Oxford University Press. Used with permission.

Graphs 2 and 3 interrelate two pairs of characters from *Pride and Prejudice* in terms of their use of the 30 most common words. The correlation between Darcy and Elizabeth (Graph 2) is strong and positive (0.930). Burrows comments:

> Few of the thirty words lie far from the diagonal line that represents parity of incidence. Among those lying to the right of the line (those for which Darcy's incidence is higher than Elizabeth's), 'was' offers the most marked divergence. (1987: 82)

This is due to his use of the past tense in his long letter of explanation. Burrows continues his commentary on the graph by pointing out that

> no one pronoun runs strongly either way: but those of the first person all lie on Darcy's side, those of the third person on Elizabeth's. On the other side of the diagonal, Elizabeth has significantly more recourse to the weakly emphatic verb-form, 'do'.

He concludes that, 'all in all, Graph [2] illustrates a suitably close resemblance between the idiolects of two strong-minded, intelligent, and essentially well-mannered characters whose disputes are conducted on even terms and whose eventual *rapprochement* is entirely credible' (Burrows 1987: 83; original emphasis). In contrast, we see a far weaker positive correlation between Collins and Lydia in Graph 3, suggesting that they speak strongly divergent idiolects. Burrows comments on the value in comparing correlation coefficients: 'To compare Graphs [2] and [3] is, in short, to see a strong contrast between an essentially similar and rather dissimilar pair of idiolects and to realize the immense difference between 0.930 and 0.598 as correlation-coefficients' (1987: 85).

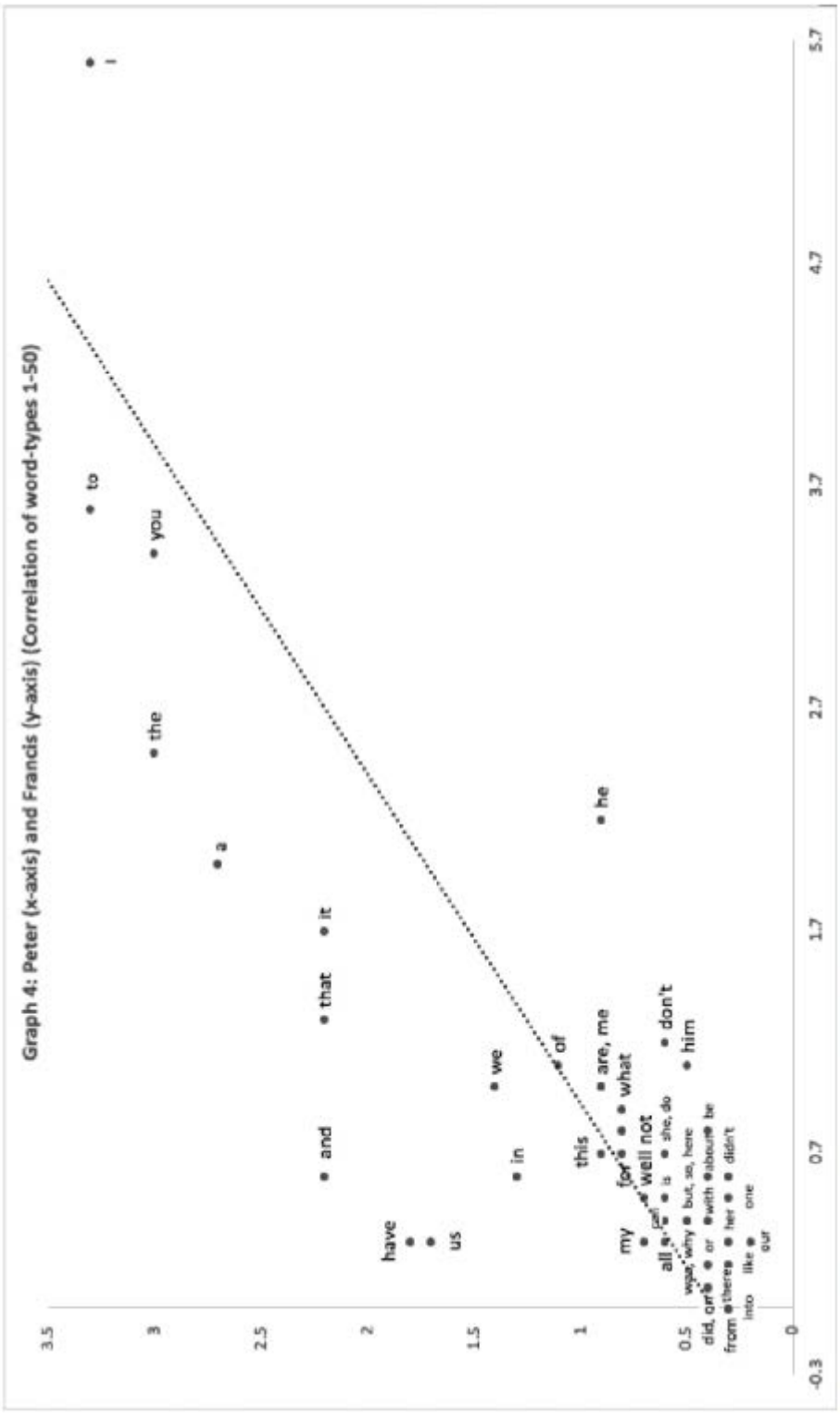
Linear regression and the correlation coefficient are also applicable to screenplay dialogue. The textalyser software generates a word list and their frequency. From the dialogue of *The Darjeeling Limited* I have collected from the analysis of each character's dialogue the 50 most common words they share. I then used Excel to calculate the correlation coefficient between the vocabulary of the three brothers (their use of the same 50 most common words):

Correlation between Francis and Jack: 0.741

Correlation between Francis and Peter: 0.803

Correlation between Peter and Jack: 0.884

Each one can be represented in a linear regression graph comprising all the words. Graph 4 presents the comparison between Francis and Peter, which I have plotted using Excel (see Graph 4).



Graph 4: Peter (x-axis) and Francis (y-axis) (correlation of word-types 1–50).

Graph 4: Peter (x-axis) and Francis (y-axis) (Correlation of word-types 1–50).

The x-axis on the bottom of Graph 4 represents the average of Peter's use of the most common words and the y-axis represents the average of Francis's use of the most common words (expressed as percentages). If a word ends up on the regression line (such as 'of') this means it is used equally by both brothers. Words above the line are used more frequently by Francis than by Peter and words below the line are used more frequently by Peter than by Francis.

There is a strong positive correlation between the two brothers' vocabulary (0.803), although this is not as strong as the correlation between Peter and Jack (0.884). In terms of pronouns, 'I', 'he' and 'our' are on Peter's side of the line, while 'you' and 'us' are on Francis's side (confirming the comments I made earlier). In addition, the frequency of use of function words can tell us about the grammatical structure of their speech, which link up to data already collected on sentence length and readability. This stylometric profile of screenplay characters makes possible very accurate comparisons between them. In particular, we can discover through quantitative methods if all the characters sound the same by sharing the same idiolect (vocabulary, grammar) or if the screenwriter is able to differentiate between characters in terms of vocabulary and grammar. (Burrows argues that Jane Austen's literary imagination enabled her to do this with great precision.) In future research I will need to contextualize the figures in Tables 3–5 and Graph 4 by conducting extensive comparative stylometric studies of screenplays, beginning with Wes Anderson's other work.

Conclusion

Is stylometry of value when applied to screenplays? At the beginning I pointed out that this article is exploratory and primarily illustrative to the extent that my aim is to try out on one Wes Anderson film the statistical methods that Burrows applied to Jane Austen's dialogue. There are at least four reasons why stylometry can be applied to screenplays:

First, simple descriptive statistics can represent the screenplay economically: Word frequency counts not only inform us of the amount of words each character speaks in relation to each other, but it can also inform us of the average length of the sentences they speak, the average syllables per word, the readability of the dialogue, the balance between function and lexical words, use of pronouns, typical sentence beginnings, how many sentences are questions, etc.

Second, using time series analysis (see Burrows: 1987, Chapter 10), stylometry can quantify how dialogue changes when characters talk to different characters, or more generally it can chart the changes in a character's speech across the entire screenplay.

Third, for co-written screenplays, we can use stylometry for authorship attribution to determine who wrote which parts of the screenplay. (*The Darjeeling Limited* has three writing credits and three brothers – is there a correlation between writers and characters? Can we attribute the differences between characters to different writers?)

Fourth, one can compare and contrast characters from different films written by the same screenwriter to determine if there are linguistic similarities between them (Owen Wilson's characters in Wes Anderson's films, for example – are they all variations of

the same character?). This links up with the narratological concept of storyworld, an abstract totality encompassing everything that fictionally exists across a director's films, in which each film is simply the partial manifestation of an author's universe. Whereas in previous research (Buckland 2019) I examined Wes Anderson's storyworld in terms of abstract codes and structures (paradigms, kinship structures, binary oppositions, mediators, systems of exchange and rules of transformation), I now plan to examine Wes Anderson's storyworld again via a stylometric analysis of his screenplays inspired by the work of John Burrows. More generally, I aim to use the work of Burrows as the starting point in establishing a new research programme within screenplay studies, one based on the stylometric analysis of the language of screenplays.

References

Buckland, Warren (2019), *Wes Anderson's Symbolic Storyworld: A Semiotic Analysis*, New York: Bloomsbury.

Burrows, John F. (1987), *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*, Oxford: Clarendon Press.

Collins, Wilkie (1868), *The Moonstone*, London: Tinsley Brothers.

Craig, Hugh and Greatley-Hirsch, Brett (2017), *Style, Computers, and Early Modern Drama: Beyond Authorship*, Cambridge: Cambridge University Press.

Ginzburg, Carlo and Davin, Anna (1980), ‘Morelli, Freud and Sherlock Holmes: Clues and scientific method’, *History Workshop*, volume 9, issue 1, pp. 5–36.

Hoover, David L., Culpeper, Jonathan, and O’Halloran, Kieran (2014), *Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama*, New York: Routledge.

McNamara, Danielle S., Graesser, Arthur C., McCarthy, Philip M. and Cai, Zhiqiang (2014), *Automated Evaluation of Text and Discourse with Coh-Metrix*, New York: Cambridge University Press.

Nelmes, Jill (2011), ‘Realism and screenplay dialogue’, in J. Nelmes (ed.), *Analysing the Screenplay*, London: Routledge, pp. 217–36.

Piazza, Roberta, Bednarek, Monika and Rossi, Fabio (eds) (2011), *Telecinematic Discourse: Approaches to the Language of Films and Television Series*, Amsterdam: John Benjamins.

Price, Steven (2010), *The Screenplay: Authorship, Theory and Criticism*, Basingstoke: Palgrave-Macmillan.

Rabiger, Michael (2003), *Directing: Film Techniques and Aesthetics*, 3rd ed., Amsterdam: Focal Press.

Rockwell, Geoffrey and Sinclair, Stéfan (2016), *Hermeneutica: Computer-Assisted Interpretation in the Humanities*, Cambridge, MA: MIT Press.

Webster Foster, Hannah (1797), *The Coquette,* Boston: William Fetridge.

Wollheim, Richard (1973), 'Giovanni Morelli and the origins of scientific connoisseurship', *On Art and the Mind: Essays and Lectures*, London: Allen Lane, pp. 177–201.

Contributor details

Warren Buckland is Reader in Film Studies at Oxford Brookes University, UK. His recent publications include *Wes Anderson's Symbolic Storyworld* (2019), *Hollywood Puzzle Films* (ed. 2014), *Film Theory: Rational Reconstructions* (2012), and *Puzzle Films: Complex Storytelling in Contemporary Cinema* (ed. 2009).

Contact:

Warren Buckland
School of Arts
Oxford Brookes University
Headington Hill
Oxford
OX3 0BP
UK.
wbuckland@brookes.ac.uk