Register variation explains stylometric authorship analysis

Grieve, Jack

DOI:
10.1515/cllt-2022-0040

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Grieve, J 2023, 'Register variation explains stylometric authorship analysis', *Corpus Linguistics and Linguistic Theory*, vol. 0, no. 0. https://doi.org/10.1515/cllt-2022-0040

Article

Jack Grieve*

Register variation explains stylometric authorship analysis

https://doi.org/10.1515/cllt-2022-0040
Received May 6, 2022; accepted November 3, 2022; published online January 2, 2023

Abstract: For centuries, investigations of disputed authorship have shown that people have unique styles of writing. Given sufficient data, it is generally possible to distinguish between the writings of a small group of authors, for example, through the multivariate analysis of the relative frequencies of common function words. There is, however, no accepted explanation for why this type of *stylometric* analysis is successful. Authorship analysts often argue that authors write in subtly different dialects, but the analysis of individual words is not licensed by standard theories of sociolinguistic variation. Alternatively, stylometric analysis is consistent with standard theories of register variation. In this paper, I argue that stylometric methods work because authors write in subtly different registers. To support this claim, I present the results of parallel stylometric and multidimensional register analyses of a corpus of newspaper articles written by two columnists. I demonstrate that both analyses not only distinguish between these authors but identify the same underlying patterns of linguistic variation. I therefore propose that register variation, as opposed to dialect variation, provides a basis for explaining these differences and for explaining stylometric analyses of authorship more generally.

Keywords: forensic linguistics; idiolect; language variation and change; multidimensional analysis; variationist sociolinguistics

1 Introduction

The analysis of disputed authorship has provided considerable evidence that everyone has a unique style of writing. Given sufficient amounts of data, it is

***Corresponding author: Jack Grieve**, Department of English Language and Linguistics, University of Birmingham, Birmingham, UK; and Alan Turing Institute, London, UK,
E-mail: j.grieve@bham.ac.uk

generally possible to distinguish between texts written by a small group of authors with a very high degree of accuracy through the quantitative analysis of a variety of textual measurements (Grieve 2007; Koppel et al. 2013). This task is referred to as *authorship attribution* and the use of quantitative linguistic analysis to resolve this task is referred to as *stylometry*. Most commonly, stylometry involves the multivariate statistical analysis of the relative frequencies of common words, especially function words (Argamon 2018; Binongo 2003; Burrows 2002; Grieve 2007; Stamatatos 2009). Stylometric analysis has notably been applied across a wide range of disciplines with great success. For example, in literary analysis, the *New Oxford Shakespeare* uses stylometric evidence to attribute plays, acts, and scenes (e.g. Taylor and Egan 2017; Taylor et al. 2016), while in forensic analysis, stylometric methods are increasingly employed to help resolve cases of disputed authorship in support of the delivery of justice (e.g. Grieve and Woodfield 2021; Juola 2012; Kredens et al. 2019).

Although research in stylometry has focused on how to resolve cases of disputed authorship accurately, it is important to understand why these methods are successful. For example, why do people differ in terms of their rates of usage of different function words? What is the theoretical basis for methods that exploit these patterns of linguistic variation? Such questions matter for several reasons. In a forensic context, a theoretical basis is often a legal requirement for the application of any type of scientific analysis as a source of evidence. For example, in the United States, the Daubert Standard requires that expert testimony be grounded in generally accepted scientific principles (Daubert v. Merrell Dow Pharmaceutical 1993) (see Groscup et al. 2002). More generally, understanding how any method works can help us maximise its performance, for example, by identifying areas for refinement or likely sources of bias. Finally, the distinctiveness of individual style, which is a very robust empirical finding supported by over a century of research in stylometry (see Grieve 2005; Stamatatos 2009), is a phenomenon which theories of language variation and change should be expected to explain.

Remarkably, however, the scientific basis for stylometry is a topic that has received very little attention in applied or theoretical linguistics (although, for discussions, see Grant and Baker 2001; Kestemont 2014; Nini 2013, 2023; Nini and Grant 2013; Wright 2017), even though these techniques tend to be based on relatively simple and transparent statistical methods, allowing for the patterns of language variation that underlie their successful application to be directly observed. At least in part, this situation reflects the fact that stylometric findings can be difficult to reconcile with standard theories of language variation and change, and that much of the relevant research has been conducted outside of linguistics – in literature, statistics, and computer science (for a history of the field,

see Grieve 2005; Stamatatos 2009). The empirical results of stylometry therefore provide a theoretical challenge for linguists, just as much as they do for authorship analysts.

Of course, authorship analysts have proposed explanations why, for example, the relative frequencies of function words vary across authors, but these explanations are generally inadequate. Perhaps most commonly, these proposals highlight that these features are frequent and primarily reflect the grammatical structure, as opposed to the topical content, of a text. These are not really explanations, however, only further descriptions of the phenomenon to be explained. Analysts have also claimed that the use of high-frequency grammatical features is subconscious. This assumption may be true for spontaneous speech, but it also seems, in a sense, to be true for the use of most content words in such contexts, and to be false for the use of most function words in carefully edited texts, as are often considered in stylometry, where each word has likely been read many times over. Furthermore, even if true, this claim once again does not actually provide a direct explanation for why such patterns can distinguish between the writings of different authors: subconscious behaviours are not necessarily unique.

A more productive approach to explaining the distinctiveness of individual style is to draw directly on theories of language variation. Variationist theories of sociolinguistics (e.g. Eckert 2012; Labov 1972; Tagliamonte 2011) are probably most commonly evoked by authorship analysts. The basic claim is that we know language varies in systematic ways depending on the social and regional background of a person, and that this type of *dialect variation* is *hierarchical*. For example, we can distinguish between dialects at the levels of nations, regions, cities, and neighbourhoods, with each more narrowly defined dialect characterised by more narrowly defined patterns of linguistic variation. By extension, it seems reasonable to assume that dialect variation can be defined down to the level of the individual. All people have slightly different social backgrounds and identities and all people should therefore be expected to use slightly different dialects, including when writing, as we know dialect features extend even to standard forms of written language (Grieve 2016). These individual dialects are often referred to as *idiolects* (Hockett 1958).

It might therefore seem that a sociolinguistic theory of idiolect provides a basis for authorship analysis, as forensic linguists and other analysts have often claimed (e.g. Coulthard 2004). This, however, is only true to the extent that the analysis of authorship is consistent with the theoretical and methodological assumptions of sociolinguistics, including, most notably, the types of linguistic variables that are analysed. Sociolinguists generally focus on what are known as *sociolinguistic variables* (or *alternation variables*), which consist of sets of distinct linguistic forms

(or *variants*) with equivalent *referential* (or *denotational*) meaning – essentially alternative ways of saying the same thing (Labov 1972; Pijpops 2020; Tagliamonte 2011). Variants can be defined at any level of linguistic analysis. For example, the choice between the pronunciations of the vowel in the word *bath* is a phonological alternation, the choice between the past-tense forms *swam* and *swimmed* is a morphological alternation, and the choice between *trunk* and *boot* is a lexical alternation. Crucially, the analysis of sociolinguistic variation is generally expected to conform to the *principle of accountability* (Labov 1972), which stipulates that, when analysing a sociolinguistic variable, the full set of attested variants should be taken into consideration. The goal is to control for variation in referential meaning so as to focus on variation in structure. For example, a sociolinguistic analysis would not consider the frequency of the word *swam* in a corpus, but rather the frequency of the word *swam* measured relative to the frequency of the word *swimmed*. This is a basic assumption of sociolinguistics, inherited directly from dialectology, and adopted by many other branches of research on language variation and change.

This is also an assumption that is sometimes adopted in authorship analysis, especially in a forensic context, and especially when a *stylistic* approach, as opposed to a *stylometric* approach, is adopted. A stylistic approach involves the identification of a bespoke set of distinctive features through a close reading of the textual evidence (Coulthard 2004; Grant 2022). For example, in the Jenny Nichols case, Malcolm Coulthard considered lexical, grammatical, and orthographical alternations in text messages sent from the teenager's phone, which were suspected to have been sent by her adult lover, David Hodgson (including *my/my, to u/2u*, and *isn't*/*aint*), showing that the messages were more consistent with Hodgson's style (Coulthard et al. 2016). Similarly, Gerald McMenanin's book on forensic authorship analysis (McMenamin 2002) focuses on these types of alternations, explicitly grounding his approach in the theoretical assumptions of variationist sociolinguistics (see also McMenamin 2010). More recently, forensic linguists (e.g. Grant and MacLeod 2018, 2020) have begun to draw on *third wave* theories of sociolinguistic variation (see Bucholtz and Hall 2004; Eckert 2012) as a basis for authorship analysis, where linguistic variation is seen as a resource that individuals use for the construction of identities.

Although sociolinguistic theory may therefore provide a foundation for research in forensic stylistics, the problem for stylometry is that it is not consistent with the assumptions of variationist sociolinguistics. Rather than focus on the analysis of alternations, where the frequencies of forms are measured relative to other equivalent variants across the texts in the corpus, stylometry has focused on the analysis of the frequencies of large numbers of individual forms, such as common function words, measured relative to the total number of words across

the texts in the corpus. Patterns in these types of variables cannot be directly accounted for by standard sociolinguistic theory (for discussion, see Grieve et al. 2017). In most cases, the individual features cannot even be considered variants of any putative alternation. What words are in direct alternation with *the* or *it* or *of*? The analysis of the frequencies of individual forms directly violates the principle of accountability and these types of linguistic variables are therefore often explicitly *excluded* as valid measures in sociolinguistic analysis. For example, Dennis Preston wrote that the relative frequencies of individual forms "do not meet the basic requirements for the study of variation—the choice of more than one semantically equivalent element in environments where all have a privilege of occurrence" (Preston 2001: 291). The empirical data amassed through stylometry, as well as other fields of linguistic analysis, provide indisputable empirical evidence that such claims are false, but it is clear, nevertheless, that variationist sociolinguistics – and, by extension, the *sociolinguistic conception of the idiolect* – cannot provide a theoretical basis for stylometry.

The question I attempt to answer in this paper is therefore *what is the theoretical basis for standard stylometric methods for authorship attribution?*

A path forward is provided by Bernard Bloch's original definition of *idiolect*, where he defined the idiolect not simply as an individual's dialect but as "the totality of possible utterances of one speaker at one time in using a language to interact with one other speaker" (Bloch 1948: 7). Bloch goes on to explain that he includes the qualification "at one time" so as to "provide for the fact that a speaker's manner of speaking changes during his lifetime" and that he includes the qualification "with one other speaker" to "exclude the possibility that an idiolect might embrace more than one style of speaking." When he introduced the term *idiolect* to linguistics, Bloch therefore very much appreciated the issue of *situational variation*: a person's style does not only depend on their social background, but on the ever-changing situations in which they communicate, including the nature of their audience.

The idea that the structure of language varies depending on situational factors, i.e. across *communicative contexts*, is also the focus of research on *register variation* (see Biber and Conrad 2019; Halliday 1978). Research in this tradition has shown that the use of individual grammatical forms varies systematically depending on a wide range of situational factors – including not only audience, but modality, medium, topic, and setting. Furthermore, this research has shown that this variation directly reflects the varying communicative constraints and affordances associated with different contexts, as well as the varying goals of people who communicate in these contexts. For example, people tend to use pronouns more often in face-to-face conversations compared to many other situations because interlocutors share the same immediate visual frame of

reference, while people tend to use past tense verbs more often when telling narratives because they are recounting events that took place in the past. Notably, these types of *functional* explanations are generally rejected in variationist sociolinguistics, at least as general explanations for language change (see Labov 2001: Ch. 19).

One form of register analysis, known as *multidimensional register analysis* (see Biber 1988, 1995; Biber and Conrad 2019), is especially relevant to stylometry, as it is based on very similar methods. Like stylometry, the relative frequencies of individual grammatical forms are measured across texts in a corpus – in this case, representing different registers (e.g. conversation, reportage, fiction) as opposed to different authors. Furthermore, like stylometry, these measurements are then subjected to multivariate statistical analysis to identify aggregated dimensions of linguistic variation that differentiate between the texts in the corpus. Unlike stylometry, however, these patterns are then *interpreted* functionally as dimensions of register variation, based on the linguistic features most strongly associated with these dimensions and the texts in which these linguistic features most commonly occur.

My central claim in this paper is that the study of register variation, especially as pursued in the tradition of multidimensional register analysis, provides a theoretical basis for explaining stylometry, both for the application of these methods in general and for the findings of individual studies. In other words, rather than explaining stylometric results as a form of *individual dialect variation*, I argue that stylometric results are a form of *individual register variation*. Stylometric methods do not work primarily because all authors use distinct dialects of the same language – although this may well be the case – but because all authors use slightly different registers of the same language in a given communicative context. My goal is not to call into question the existence of the idiolect or the existence of a sociolinguistic component of the idiolect, but to recognise the situational component of the idiolect, as Bloch did when he defined the term, and to argue that it is this type of individual register variation, as opposed to individual dialect variation, that generally underlies the successful applications of standard stylometric methods for authorship analysis.

To support this claim, I present parallel stylometry and register analyses of the same corpus, which consists of a large number of opinion articles by two authors who come from very similar social backgrounds and who are writing in very similar communicative contexts. I show that both approaches not only distinguish clearly between the writings of these two authors but identify nearly identical dimensions of linguistic variation. Furthermore, I argue that the distinctive dimension identified by the register analysis can be interpreted functionally and that this same explanation accounts for the distinctive dimension

identified by the stylometric analysis. In this way, I show that a register analysis can provide a basis for explaining the results of a stylometric analysis. Finally, I conclude by considering what these results might reveal about the nature of linguistic variation and change more generally.

2 Data

In this study, I compare two corpora of 130 newspaper articles written by two prominent British opinion columnists, both of whom wrote for the *Daily Telegraph*: William Hague and Charles Moore. I collected this dataset in May 2019 using Nexis, searching backwards and downloading the first 130 opinion articles by these two authors that were over 500 words long, after removing text outside the main body of each article. The two sub-corpora are intended to represent the varieties of language used by each author in their *Daily Telegraph* opinion articles in the late 2010s.

I selected these two columnists for analysis primarily because a relatively large number of comparable articles were easily accessible for each. Furthermore, in addition to both writing weekly opinion-based political columns from a conservative standpoint for the same newspaper at the time of data collection, both authors have similar social backgrounds, broadly speaking – both being white, upper-class, British males of around the same age, who were educated in the Humanities at Oxford and Cambridge. It is also notable that both now sit in the House of Lords. Hague was born 1961 in Yorkshire and was a Conservative Member of Parliament from 1989 to 2015, including acting as the leader of the Conservative Party from 1997 to 2001. Following his retirement in 2015, he was awarded a life peerage, and began writing a weekly opinion column for the *Telegraph*, before moving to *The Times* in 2021. Charles Moore was born in 1956 in Sussex and joined the *Daily Telegraph* Newspaper in 1979 as a political columnist, being promoted to Editor of *The Spectator* in 1984, *The Sunday Telegraph* in 1992, and *The Daily Telegraph* in 1995, stepping down in 2003 to focus on writing, including authoring columns for the *Daily Telegraph*. He was awarded a life peerage in 2021.

In total, the corpora contain 284,672 words across the 260 texts, with 143,073 words for Hague and 142,599 words for Moore. The median length of Hague's texts is 1,098 words, with half his texts containing between 1,080 and 1,120 words. His shortest text contains 582 words, while his longest contains 2,277 words, although his second longest text contains 1,212 words. Hague's articles are therefore characterised by a high level of consistency in text length. The median length of Moore's texts is 1,237 words, with half his texts containing

between 916 and 1,261 words. His shortest text contains 710 words, while his longest contains 1,360 words. Moore's articles are therefore characterised by somewhat greater variation in text length. The two sub-corpora also span similar date ranges: Hague's articles were published between May 2016 and April 2019, while Moore's articles were published between October 2017 and May 2019.

Overall, the two authors' sub-corpora are therefore comparable in terms of text length, corpus size, register, and time period. The corpus also contains a relatively large amount of data for each author. This would be ideal for a corpus of possible authors compiled to help resolve a real case of disputed authorship. For example, given an anonymous opinion article believed to have been written by one of these two authors during this period, attribution would ideally be based on a corpus of a relatively large number of opinion articles written by each author from this period. Although comparing the style of two authors writing in such similar registers makes identifying distinctive features more difficult, it increases the likelihood that real differences are identified between the styles of these authors when writing in one narrowly defined communicative context.

3 Analysis

Using this corpus, I conducted three analyses. I first conducted a standard stylometric analysis based on the relative frequencies of the 50 most common function words in the corpus, which I subjected to a principal component analysis. I then conducted a standard multidimensional register analysis based on the relative frequencies of 66 grammatical features, which I subjected to a factor analysis. Finally, I compared the results of these two analyses to see if the register analysis could provide a basis for explaining the results of the stylometric analysis.

3.1 Stylometric analysis

In this section, I present the results of a standard stylometric analysis of the authorship corpus following the basic method outlined in Binongo (2003), which I will refer to as a function word principal component analysis (FW-PCA). My immediate goal is to assess whether FW-PCA can be used to successfully distinguish between the writing style of these two authors.

The FW-PCA approach is one of the two most common methods for multivariate stylometric analysis, along with Burrow's Delta (Burrows 2002). For example, these are the two main methods made available by the popular "stylo"

library in R (Eder et al. 2016). Both methods are similar from a technical standpoint: they are multivariate methods for dimension reduction, which, when given a set of quantitative linguistic variables measured across a set of texts, project the texts into low-dimensional space so that the questioned document can be attributed. The main difference between FW-PCA and Delta is the statistical method used for dimension reduction, although the two methods are also often used to analyse somewhat different linguistic feature sets. I chose to apply FW-PCA rather than Delta in this study primarily because FW-PCA allows the linguistic features associated with each aggregated dimension to be scrutinised directly, facilitating interpretation. This type of information is not directly accessible when using Delta because data is aggregated by measuring the distances between texts based on the full set of variables. Despite this difference, when applied to the same dataset, these two approaches tend to yield similar results

Specifically, I used FW-PCA to analyse the relative frequencies of the top 50 function words in the corpus, broadly following the process for feature extraction described in Binongo (2003). First, a list of the most frequent word forms (i.e. case-insensitive strings of alphabetic characters) was extracted from the full corpus. Second, the 50 most common function words were extracted from this list (two content words, *EU* and *people*, which were among the top 50 most common words, were excluded from the analysis). These function words are from any word class except nouns, adjectives, lexical adverbs, and lexical verbs, but do include auxiliary verbs, modal verbs, and adverbs of degree and negation. A number of these words can be assigned to multiple word classes. The most common function word in the corpus is *the*, which occurs with a median frequency of 65 times per article. All 50 function words occur at least once on average per article, with *her* and *she* being the two least common of the top 50 function words in the corpus, each occurring on average once per article. The complete list is provided in Figure 1. Third, to control for variation in text length, the relative frequencies of these 50 function words were calculated per thousand words across the 260 texts in the corpus, yielding a 50 function word by 260 text relative frequency data matrix.

Before discussing the multivariate analysis of the full data matrix, it is notable that some of these features show clear differences between the two authors individually. For example, Figure 2 presents individual kernel density estimation plots for the relative frequency distributions of 10 function words between the two authors, selected to illustrate a range of feature types and patterns of variation. In each case the range of observed relative frequencies of the feature across texts in the corpus is represented by the horizontal axis and the number of texts in the corpus that exhibit relative frequencies in that range is

A	About	All	An	And
Are	As	At	Be	But
By	Can	For	From	Has
Have	He	Her	His	I
If	In	Is	It	More
No	Not	Of	On	One
Or	Our	S	She	So
That	The	Their	There	They
This	To	Was	We	What
Which	Who	Will	With	Would

Figure 1: FW-PCA: Feature set.

represented by the vertical axis, split by author, with Hague in blue and Moore in red. Based on these graphs, we can see, for example, that Hague uses the forms *to, and,* and *be* substantially more often than Moore, whereas Moore uses the forms *was*, *I,* and *it* substantially more often than Hague.

To identify common patterns of variation in function frequencies, I subjected the data matrix to principal component analysis (PCA), a well-established method for dimension reduction. FW-PCA begins with the computation of a *correlation matrix* – the correlation between every pair of linguistic variables in the data matrix across the texts in the corpus. Based on this correlation matrix, a series of aggregated dimensions are then extracted, representing the most important independent patterns of linguistic variation in the underlying data matrix, where each dimension accounts for a decreasing amount of variance. Each dimension is associated with *dimension loadings*, whose magnitudes specify how strongly each linguistic variable is represented by that dimension, and whose signs specify which variables are positively and negatively correlated with each other: variables that are positively correlated are assigned the same sign, whereas variables that are negatively correlated are assigned opposing signs. Each dimension is also associated with *dimension scores*, specifying which

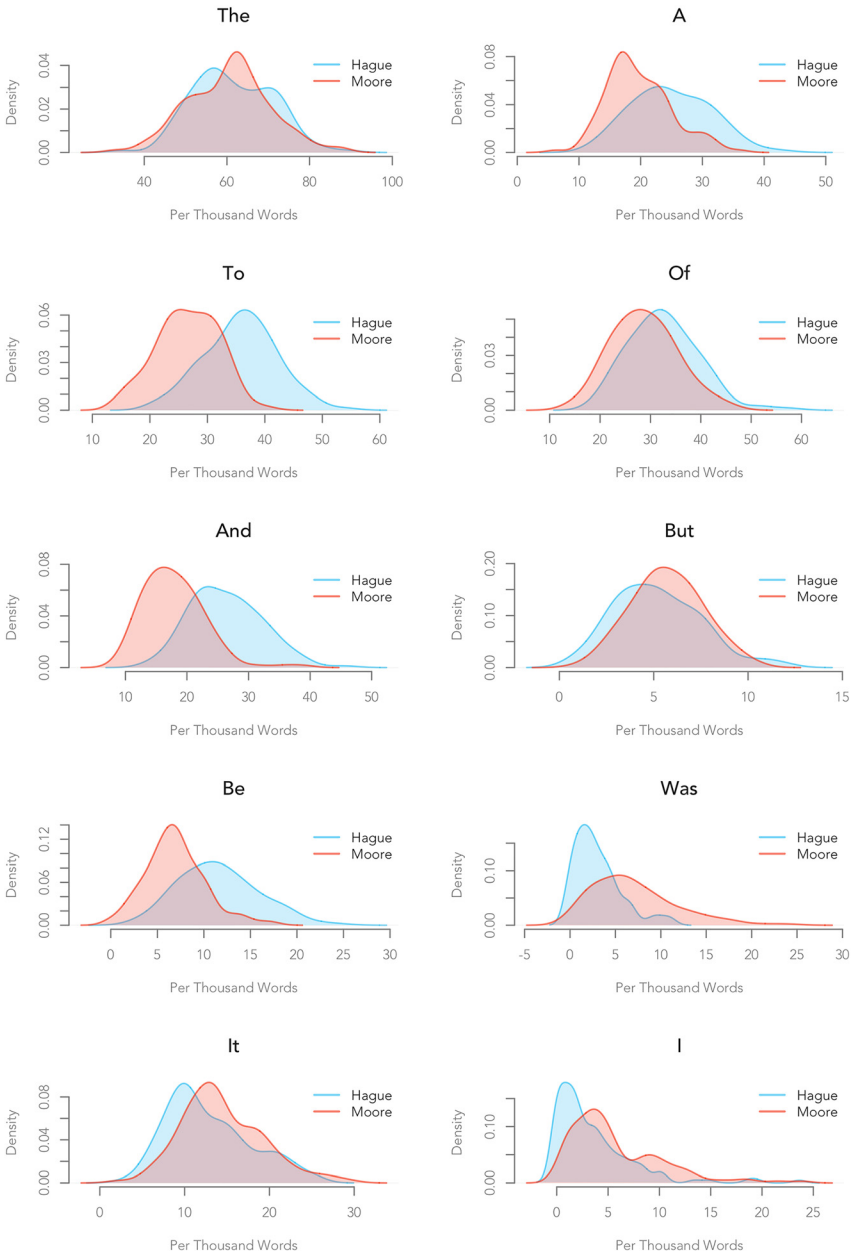

Figure 2: FW-PCA: Individual variable comparison.

texts are characterised by similar patterns of linguistic variation, where texts assigned strong positive scores are characterised by frequent use of variables assigned positive loadings and infrequent use of variables assigned negative loadings, and where texts assigned strong negative scores are characterised by frequent use of variables assigned negative loadings and infrequent use of variables assigned positive loadings.

In this way, a data matrix where each text is defined by a large number of individual linguistic variables is reduced to a data matrix where each text is defined by a smaller and more informative set of aggregated variables, each accounting for as much variation in the original data matrix as possible. No distinction is made between texts written by different authors up to this stage of the analysis, but once the dimensions are extracted, the texts written by the authors are compared based on the dimension scores so as to identify any dimensions that distinguish between their writings with a reasonable degree of accuracy. Because these dimensions are independent of each other, usually only one dimension distinguishes strongly between the authors, especially when only two authors are considered. Finally, the disputed text is projected onto the distinctive dimension and attributed to the author within whose range it falls. In this case, however, because my goal is simply to identify any dimension that distinguishes between the writings of Hague and Moore, rather than attribute a questioned document, this final step is omitted.

The scores for the 260 texts on the first and second dimensions are presented as a scatter plot in Figure 3. The FW-PCA identifies clear differences between the style of writing used by Hague and Moore. Dimension 1 shows an especially strong difference, with the vast majority of Hague's texts being assigned positive Dimension 1 scores (94.6%), and the vast majority of Moore's texts being assigned negative Dimension 1 scores (91.5%). The Dimension 1 scores are also compared directly between the two authors in the first kernel density estimation plot in Figure 4, once again showing a clear difference, with relatively little overlap. Dimension 2 also shows a weak difference between the two authors, with a slightly higher proportion of Hague's texts being assigned negative Dimension 2 scores than Hague, as can be seen in the second kernel density estimation plot in Figure 4.

Overall, it is therefore clear that the FW-PCA, especially Dimension 1, successfully distinguishes between the writing styles of the two authors with a reasonable degree of accuracy. It is still unclear, however, *why* this dimension distinguishes between these two authors. How can the differences between these two styles be described and explained? To answer this question, it is necessary to propose a linguistic explanation for this dimension, a step that is almost always ignored in stylometry. I will return to this question in Section 3.3, after

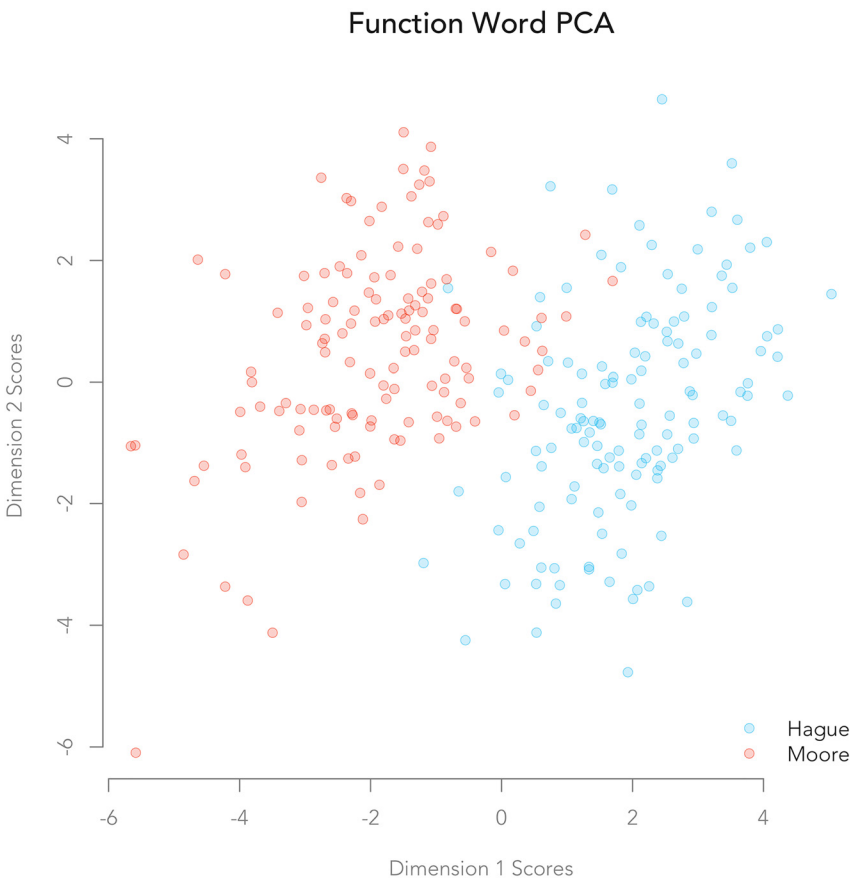

Figure 3: FW-PCA: Dimension scatterplot.

first presenting the results of a parallel register analysis of this same dataset, which also generates aggregated dimensions of linguistic variation, but which additionally requires the *interpretation* of these dimensions.

3.2 Register analysis

In this section, I present the results of a multidimensional register analysis (MDA) of the authorship dataset, following the basic method outlined in Biber (1988). My immediate goals are to assess whether MDA can be used to successfully distinguish between the writing style of these two authors and, if so, to assess

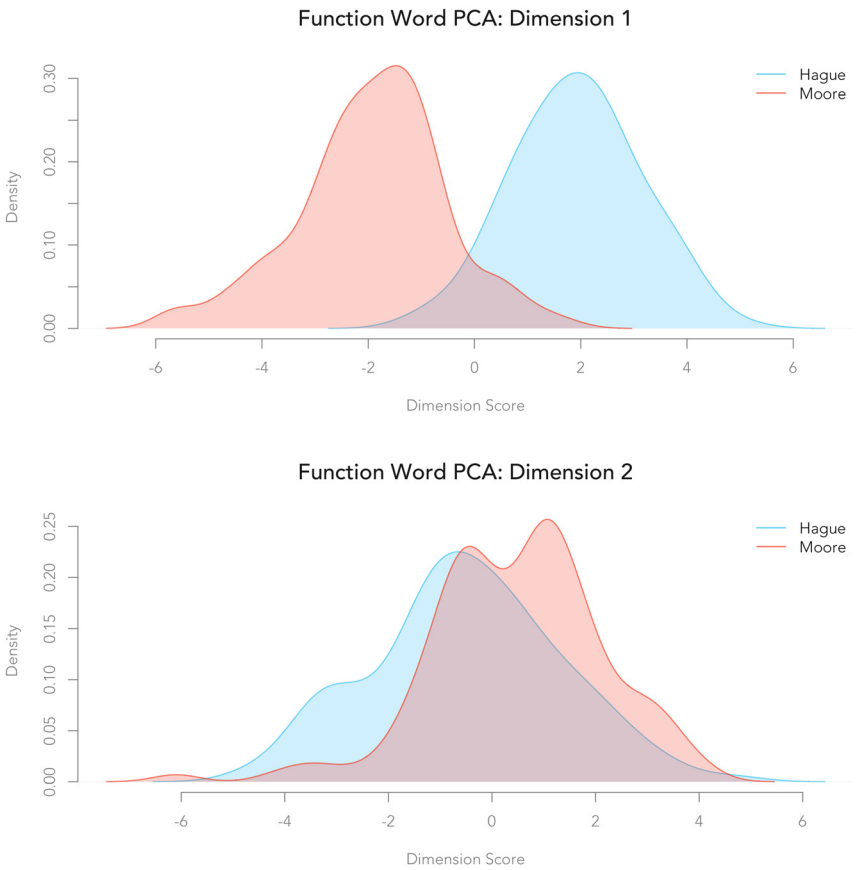

Figure 4: FW-PCA: Dimension comparison.

whether MDA can provide a theoretically grounded explanation for any variation in style that is observed.

MDA is a quantitative method for linguistic analysis developed by Douglas Biber to identify and describe underlying patterns of grammatical variation in a corpus of texts. These dimensions are then commonly used to compare the style of texts from different registers or to identify text types by clustering texts written in a similar style. The landmark study in this tradition is Biber's 1988 book, *Variation across Speech and Writing*, which presents a multivariate analysis of the relative frequencies of a diverse set of 67 grammatical features across a corpus consisting of 481 texts representing 6 spoken and 17 written registers of British English totalling 960,000 words. Based on this analysis, Biber identified and interpreted 6 dimensions of functional linguistic variation. For example, the

first dimension contrasts texts written in a more informationally dense style, especially varieties of formal writing, which tend to contain relatively large numbers of nouns and noun modifiers among other features, with texts written in a more involved style, especially varieties of spontaneous conversation, which tend to contain relatively large numbers of pronouns and verbs among other features. Alternatively, the second dimension contrasts texts written in a more narrative style, especially varieties of fiction, which tend to contain large numbers of past tense verbs and third person pronouns among other features, with texts written in a less narrative style. At a general level, this research has shown that variation in the structure of texts produced across situations is highly systematic, reflecting variation in the communicative context in which these texts are produced and the communicative purpose of the people who produce these texts.

MDA has a long history of use, having been applied to corpora representing a wide range of different languages and varieties (Sardinha and Pinto 2014, 2019), both by applying the dimensions identified in previous MDA studies (especially Biber 1988) to new corpora, and by conducting original MDA analyses, identifying new dimensions specific to the variety of language under analysis. MDA, however, has rarely been used for authorship analysis, even though it is methodologically very similar to FW-PCA – a point that has also only rarely been acknowledged in either tradition. The primary exception I am aware of is Biber and Finegan (1994), who present an MDA of texts written by four authors (Addison, Dafoe, Johnson, Swift), arguing that MDA provides a principled framework for the analysis of variation across authors grounded in a more general understanding of stylistic variation across registers (see also Nini and Grant 2013). Although, in this sense, the goals of Biber and Finegan (1994) are similar to this study, rather than conduct an original MDA, they reused three dimensions from Biber (1988). Consequently, their ability to distinguish between the authors was limited: in essence, they asked whether the general dimensions of stylistic variation identified in Biber (1988) could distinguish between these four authors, who were writing at a different time and over a more restricted range of registers, rather than if MDA more generally could distinguish between authors, as I do in this study.

In particular, I analysed the authorship corpus by calculating new dimensions of register variation based on the feature set from Biber (1988), using Andrea Nini's Multidimensional Analysis Tagger (Nini 2019), which identifies and counts the 67 features used in Biber (1988). I focused on this feature set because it provides a firm foundation for interpreting the results of MDA in a standard and accessible manner. For all 230 texts in the corpus, I measured 66 of these variables, as listed in Figure 5. All these variables can be understood as being

measured as the frequencies of linguistic forms relative to the total number of words in each text. Broadly speaking, these variables represent word classes defined at various levels of generality. I exclude type token ratio (see Baayen 2001), given issues with this measure, even when measured across equal sized samples. Figure 6 presents individual kernel density estimation plots for the relative frequencies of 10 selected features between the two authors. Based on these graphs, we can see, for example, that Hague uses nominalisations and modals of prediction substantially more often than Moore, whereas Moore uses third person pronouns and past tense verbs substantially more often than Hague.

Adjective: Attributive	Adjective: Predicative	Adverb: Amplifier	Adverb: Conjunct	Adverb: Discourse Marker	Adverb: Downtoner
Adverb: Emphatic	Adverb: Hedge	Adverb: Other	Adverbial: Place	Adverbial: Time	Adverbial Subordinate: Causative
Adverbial Subordinate: Concessive	Adverbial Subordinate: Conditional	Adverbial Subordinate: Other	Average Word Length	Contraction	Coordination: Clausal
Coordination: Phrasal	Determiner: Demonstrative	Existential There	Infinitive To	Modal: Necessity	Modal: Possibility
Modal: Prediction	Negation: Analytic	Negation: Synthetic	Noun: Gerund	Noun: Nominalisation	Noun: Other
Participial Clause: Past	Participial Clause: Present	Passive: Agentless	Passive: By-phrase	Perfect Aspect	Preposition Phrase
Preposition Stranding	Pronoun: Demonstrative	Pronoun: First Person	Pronoun: Indefinite	Pronoun: It	Pronoun: Second Person
Pronoun: Third Person	Relative Clause: Pied Piping	Relative Clause: Sentence Relative	Relative Clause: That Subject	Relative Clause: That Object	Relative Clause: WH Subject
Relative Clause: WH Object	Split Auxiliary	Split Infinitive	Subordinator That Deletion	Tense: Past	Tense: Present
That Complement: Adjective	That Complement: Verb	Verb: Be Main Verb	Verb: Seem and Appear	Verb: Private	Verb: Pro-verb Do
Verb: Public	Verb: Suasive	WH Clause	WH Question	WHIZ Deletion: Past Participial	WHIZ Deletion: Present Participial

Figure 5: MDA: Feature set.

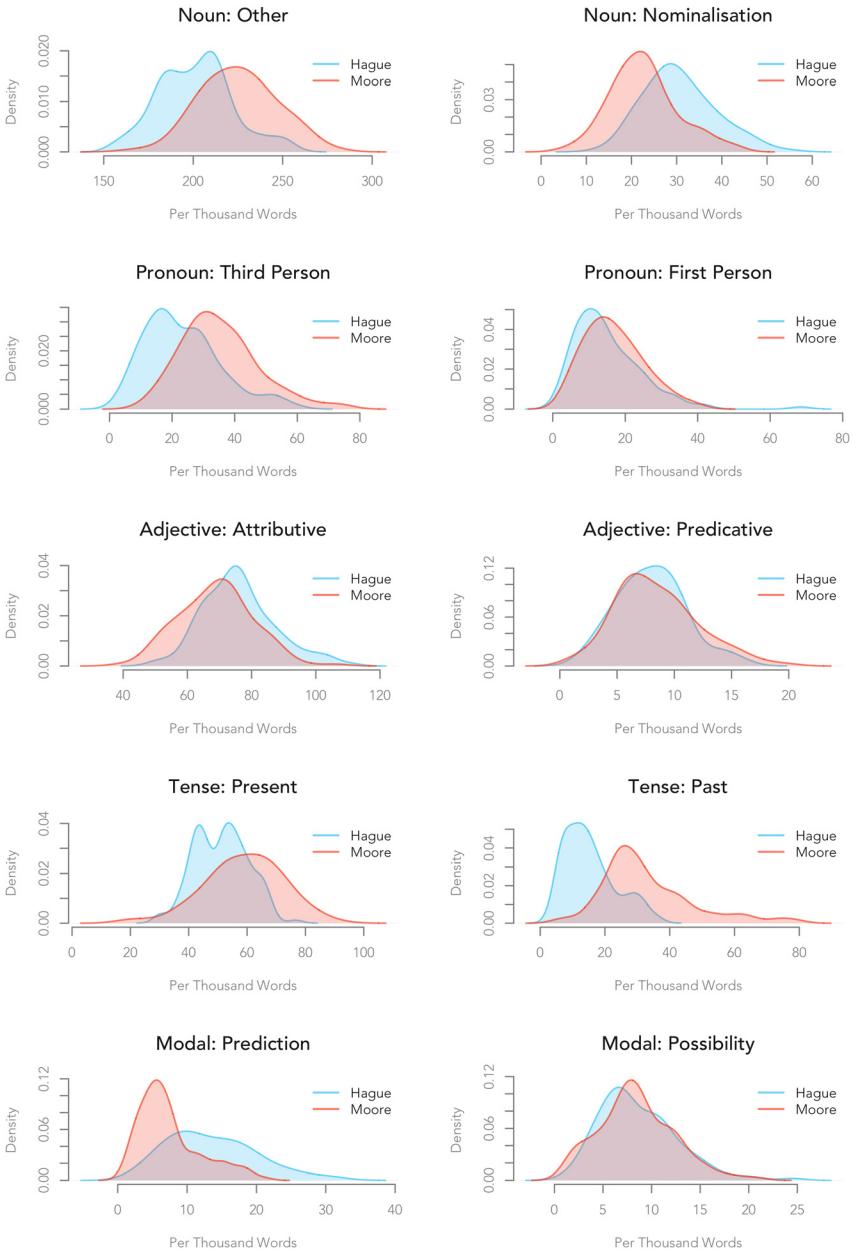

Figure 6: MDA: Individual variable comparison.

Before presenting the results of the multivariate analysis, it is important to acknowledge that the MDA feature set (see Figure 5) is related but distinct from the FW-PCA feature set (see Figure 1). At the most basic level, the FW-PCA feature set is based on word-level categories, whereas the MDA feature set is based primarily on grammatical categories, where features are computed by counting sets of grammatically related words or more complex multi-word units as opposed to individual words. Furthermore, whereas the features for FW-PCA are based on counting all tokens of a word as an instance of a given word type, features for MDA may be based on counting tokens of the same word or multi-word unit as different grammatical types (e.g. *can* as modal verb is distinguished from *can* as a noun). Despite these differences, the feature sets are similar, and in some cases, the individual features do overlap. For example, the pronoun *it* is counted as a feature in both approaches. Nevertheless, these feature sets are largely independent and represent different levels of linguistic analysis.

Next, I subjected the 66-variable-by-230-text data matrix to a multivariate statistical analysis to identify the most important dimensions of linguistic variation. This could be accomplished via PCA, as in the previous section, but MDA is generally based on Factor Analysis (FA), a related technique, which also produces loadings and scores for a series of aggregated dimensions. Outlining the technical differences between these methods is beyond the scope of this paper (see Everitt and Hothorn 2011), but PCA is generally used for simple dimension reduction, whereas FA is used for identifying and modelling latent variables: FA assumes that by inspecting correlations between observable variables, unobserved underlying variables that are responsible for these correlations can be abduced. Unlike PCA, interpreting dimensions is therefore considered a necessary step in FA. Nevertheless, in my experience, given the same dataset, stylometric and register analyses tend to produce similar results regardless of whether PCA or FA is used.

I reduced this dataset to two dimensions using FA (with promax rotation, following standard MDA procedure). As opposed to PCA, the analyst's choice of how many dimensions to consider in FA affects the composition of the other dimensions. I chose to focus on two dimensions for three reasons. First, there is a substantial drop in the amount of variance explained: extracting two dimensions accounts for 9.2 and 6.5% of the variance, whereas each additional dimension accounts for substantially less variance (less than 5% of the variance). Second, my goal is not to explore the full range of stylistic variation in this corpus, but to identify a single dimension that distinguishes between the style of the two authors. As I show, focusing on two dimensions is sufficient to meet this goal. Third, focusing on two dimensions increases the comparability of the two analyses.

The scores for the 260 texts on the first and second dimensions are presented as a scatter plot in Figure 7. The MDA identifies clear differences between the style

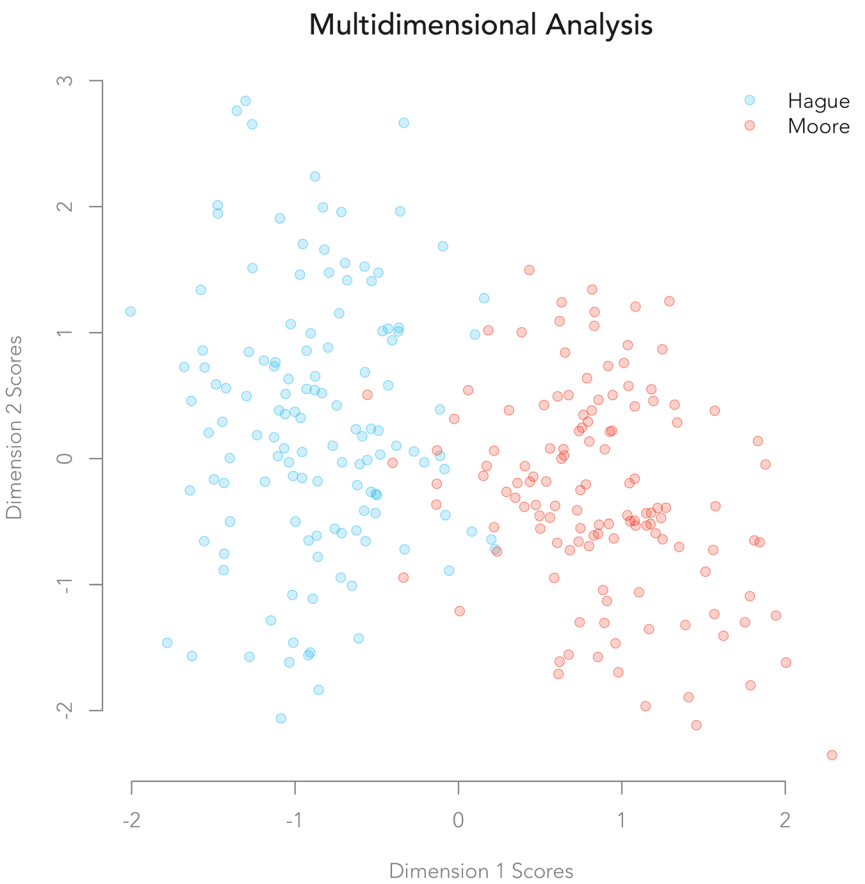

Figure 7: MDA: Dimension scatterplot.

of writing used by Hague and Moore. Dimension 1 shows an especially strong difference, with the vast majority of Hague's texts being assigned negative Dimension 1 scores (96.2%), and the vast majority of Moore's texts being assigned positive Dimension 1 scores (94.6%), slightly outperforming the FW-PCA from this perspective. The distribution of dimension scores are also compared across the two authors using kernel density estimation plots in Figure 8. Once again, Dimension 2 also shows a weak difference between the two authors, with Hague's texts being assigned slightly more positive Dimension 2 scores than Hague.

Finally, as opposed to the FW-PCA, I interpreted the results of the MDA, thereby providing a linguistic explanation for the variation observed between the two authors on MDA Dimension 1. Given the goals of this paper, I focus exclusively here on Dimension 1, as this is the dimension that distinguishes

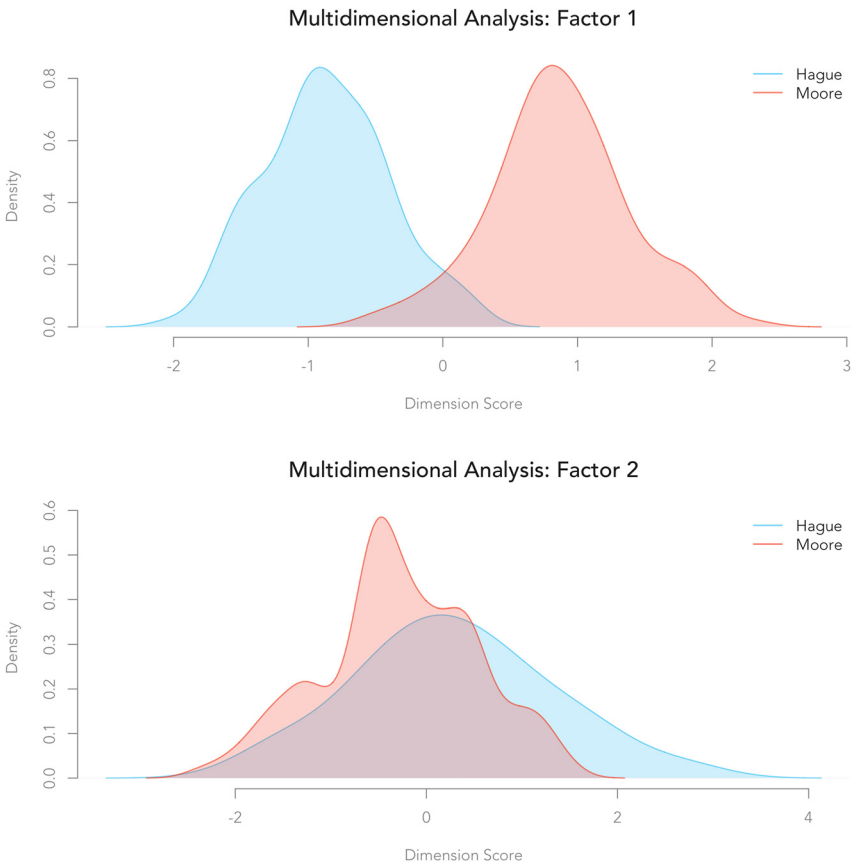

Figure 8: MDA: Dimension comparison.

strongly between these two authors. As is standard in MDA, to interpret this dimension, I considered the opposing functional characteristics of both the linguistic features (based on the dimension loadings) and the texts (based on the dimension scores) most strongly associated with either pole of this dimension. The 16 variables most strongly associated with this dimension (loadings > ±0.4) are presented in Table 1, including 8 positively loading variables, which tend to occur more often in Moore's writings and less often in Hague's writings, and 8 negatively loading variables, which tend to occur more often in Hague's writings and less often in Moore's writings. A majority of these 16 variables present a picture of linguistic variation that is well-attested in MDA research, making the interpretation of this dimension relatively straightforward.

Table 1: Register analysis: Dimension 1 loadings.

Positive features	Loading	Negative features	Loading
Tense: Past	0.78	Noun: Nominalisation	–0.64
Pronoun: Third person	0.63	Preposition phrase	–0.57
Adverbial subordinator: Causative	0.50	Infinitive *To*	–0.56
Verb: Public	0.48	Modal: Prediction	–0.54
Noun: Other	0.48	Relative clause: *That* subject	–0.48
Adverbial subordinator: Concessive	0.45	Coordination: Phrasal	–0.46
Adverb: Other	0.43	WHIZ deletion: Present participial	–0.40
Relative clause: WH object	0.40	Adjective: Attributive	–0.40

Most notably, a number of the most strongly loading positive variables are clearly related to a *narrative style*, including past tense verbs, which is the single most distinctive variable on this dimension, and third person pronouns. These are the two quintessential features of narratives, which involves recounting events in the past with recurring participants (Biber 1988). Public verbs, most notably forms of the verb *to say*, also exhibit strong positive loadings and are common in narratives, where they are generally used to report dialogue (Biber 1988). Alternatively, modals of prediction (*will, would, shall*), which are used to reference future events, show a complementary pattern, loading strongly as a negative variable. Attributive adjectives also exhibit strong negative loadings and are generally associated with non-narrative texts (Biber 1988). Dimension 1 therefore clearly identifies a distinction between Moore's use of a more narrative style of writing and Hague's use of a less narrative style of writing.

In addition, the most strongly loading negative variables, aside from modals of prediction, are all clearly associated with an *expository style*, including nominalisations and variables related to noun modification, including preposition phrases, infinitive phrases, relative clauses, WHIZ deletion, and attributive adjectives. The frequent use of phrasal coordination is also consistent with this analysis. In general, these variables are used to construct texts that maximise the amount of information incorporated into sentences by creating complex noun phrases, and have been repeatedly found to be strongly associated with informally dense registers in MDA research (Biber 1988; Biber and Conrad 2019). Notably, this style is typical of newspaper writing (Biber 1988). Alternatively, the frequent use of adverbs and adverbial subordination, which is generally associated with less informationally dense forms of communication (Biber and Gray 2010), show a complementary pattern, loading strongly as positive variables. Factor 1 therefore also clearly identifies a distinction between Hague's use

of a more expository style of writing and Moore's use of a less expository style of writing.

Overall, MDA Dimension 1 therefore identifies an opposition between a more narrative style and a more expository style, representing a combination of the first two dimensions identified in Biber (1988). This difference can be better appreciated by contrasting extracts from articles with especially strong positive and negative scores. For instance, consider Examples 1 and 2, extracted from Moore's articles with strong positive scores, which are written in a much more narrative style, with Examples 3 and 4, extracted from Hague's articles with strong negative scores, which are written in much more expository style.

1. *Then, in 2013, a victim of Smyth protested at the failure by the Titus Trust, heir to the Iwerne Trust responsible for the camps, to pursue his accusations through a proper police investigation. In early 2017, following a television exposé that door-stepped Smyth in South Africa, Archbishop Welby was criticised: he had known about the 2013 complaint but allegedly done little* (Moore, 2018-03-16).
2. *I once sat next to a Frenchwoman at dinner in Paris. "We are not anti-Semitic," she said, "It is all lies. The newspapers only say this because they are controlled by Jews." Her comically contradictory words encapsulated so much of the anti-Semitic mind–that she claimed to be falsely accused; that Jews do not tell the truth; that Jews have power* (Moore, 2018-03-30).
3. *So what do they do now? On the face of it, Theresa May and her slightly reshuffled Cabinet face nearly insurmountable constraints and dangers. The normal survival plan for a minority government is to pass little legislation, but preparation for Brexit requires a mass of complex and controversial law-making* (Hague, 2017-06-12).
4. *But even the eurozone is not the greatest threat to the unity of the EU. The crisis most likely to overwhelm Europe in the coming years and bring populist or nationalist leaders like Marine Le Pen to power is an uncontrollable rise in immigration from Africa and the Middle East. The population of these regions is expected to double over the next 30 years, which will be an increase of over a billion people* (Hague, 2017-05-08).

More generally, this opposition would appear to reflect two basic ways a columnist can express their personal opinions in this register – through the sharing of personal experiences or through logical argumentation. That is not to say that Hague and Moore write in totally different styles: both authors employ exposition and narration, and, in many ways, their articles seem more similar than dissimilar, consistent with general expectations for newspaper opinion columns. However, on the whole, Moore relies more on narrative, whereas Hague relies more on exposition. This underlying difference in style is the explanation

for how the MDA is able to distinguish between the writings of these two authors with such a high degree of accuracy based on the analysis of the relative frequencies of grammatical forms.

Notably, this interpretation does leave 2 of the 16 variables that load strongly on Factor 1 unexplained. The frequent use of WH clauses with object gaps is easily explained by Moore's more frequent use of *which* as a relative pronoun, which occurs 66 times in his corpus compared to only 9 times in Hague's, which may reflect a form of dialect variation. The frequent use of nouns, however, is more difficult to account for, especially as this is one of the most basic indicators of informationally dense texts. There is no simple resolution to this inconsistency, but it would appear to result from Hague's frequent use of other word classes to modify nouns — including prepositions, infinitives, relative pronouns, and adjectives — driving down his overall use of nouns relative to Moore. Overall, however, this one exception does not undermine the primary result of the MDA, which otherwise identifies a clear, consistent, and interpretable difference in the style of writing used by these two authors.

3.3 Comparison of results

In this section, I present a comparison of the results of the parallel stylometric and register analyses of the authorship dataset. I focus on the first dimension extracted by each analysis because they successfully distinguished between the two authors. My goals are to measure the similarity of these two dimensions and to assess if the results of the register analysis can provide a basis for explaining the results of the stylometric analysis.

Although both analyses distinguish between the articles written by the two authors with a high degree of accuracy, it is unclear how similarly the individual texts have been scored. To assess the similarity of the two dimensions, I correlated the dimensions scores across the 260 texts (i.e. as plotted on the horizontal axes of Figures 3 and 7), finding a very high negative correlation between these two dimensions ($r = -0.90$), indicating that both dimensions likely identify the same underlying pattern of variation. Note that the correlation is negative because the order of the dimensions are inverted, with Hague's texts being assigned positive scores overall on Dimension 1 of the FW-PCA, but negative scores on Dimension 1 of the MDA. The similarity of these distributions is visualised by the scatter plot in Figure 9, which graphs the FW-PCA Dimension 1 scores against the MDA Dimension 1 scores, showing this very strong negative linear relationship.

Next, I compared the variable loadings for these two dimensions to see if my interpretation of MDA Dimension 1 could provide a basis for an interpretation of

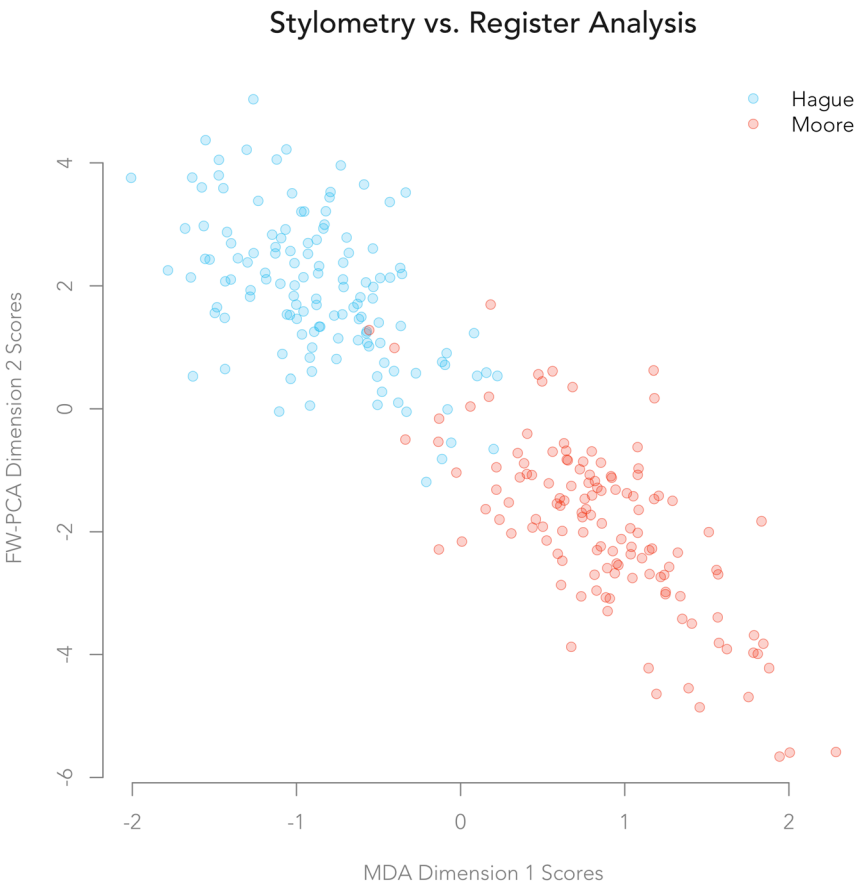

Figure 9: FW-PCA and MDA comparison.

FW-PCA Dimension 1. The loadings on Dimension 1 for the stylometric analysis are not as strong as the loadings on Dimension 1 of the register analysis; I therefore present the top 10 positive and negative loadings for the 50 function words in Table 2. Overall, it is clear that the two dimensions broadly align in terms of these feature loadings. The positive pole of MDA Dimension 1 was interpreted as being related to a narrative style, most notably loading past tense and third person pronouns. Similarly, the negative pole of FW-PCA Dimension 1 loads the past tense verb *was* and the third person pronouns *he, his, she* and *her,* as well as other pronominal forms and the relative pronoun *which,* which was identified by MDA analysis as co-occurring with these narrative features. Alternatively, the negative pole of MDA Dimension 1 was interpreted as being related

Table 2: Stylometry analysis: Dimension 1 loadings.

Positive features	Loading	Negative features	Loading
Be	0.28	*was*	−0.29
To	0.27	*he*	−0.25
Will	0.22	*s*	−0.24
And	0.21	*his*	−0.23
With	0.20	*i*	−0.18
Or	0.19	*which*	−0.17
Have	0.19	*she*	−0.14
Can	0.19	*Her*	−0.13
Would	0.18	*By*	−0.12
A	0.18	*Not*	−0.12

to an expository style, most notably loading various features related to noun modification, including prepositions and infinitives, as well as coordinating conjunctions and modals of predictions. Similarly, the positive pole of FW-PCA Dimension 1 loads various function words related to these word classes, including *to, will, and, with, or, would,* and *a*. In addition, the infinitive form of the verb *be,* which is the highest loading positive feature, also regularly co-occurs with a number of these function words.

It is therefore clear that these two dimensions not only distinguish between the two authors, but they distinguish between these authors based on *the same underlying pattern of linguistic variation*. In other words, although they draw on two different sets of linguistic features and two different types of multivariate statistical analysis, both approaches find that differences in the writing styles of these two authors is primarily driven by the same latent dimension of linguistic variation: Hauge writes in a more expository style, whereas Moore writes in a more narrative style. This kind of functional explanation is not generally provided in standard stylometric analysis, where classification accuracy is usually considered to be sufficient evidence of the validity of the method, but after conducting the MDA and comparing the results of the two analyses, it is clear that this same explanation accounts for the results of the FW-PCA.

Finally, although the goal of this analysis is not to prescribe how best to conduct authorship analysis, a brief discussion in light of these results is warranted. There are advantages and disadvantages to both approaches: FW-PCA is easier to conduct and has a longer record of successful application in this domain, whereas MDA facilitates interpretation and has a more direct theoretical foundation. Overall, however, the results of this study imply that the choice is

of no great consequence. In general, the multivariate analysis of the relative frequencies of grammatical forms can be expected to distinguish between the writings of different authors with a relatively high degree of accuracy. Personally, I will continue to use FW-PCA to help resolve cases of disputed authorship because the method is simple, replicable, and widely applied for authorship analysis. However, given the results of this study, I will now apply FW-PCA with confidence that its results can and should be explained as a form of register variation – that stylometric analysis works because authors tend to write in slightly different registers.

4 Conclusion

In this study, I have shown that articles by two authors, writing for the same newspaper and over the same period of time, can be robustly distinguished through the stylometric analysis *because* they wrote in subtly different registers: one author wrote in a more narrative style, while the other author wrote in a more expository style. In other words, I have argued that there is a functional explanation for the differences observed in the writing styles of these two authors: these two authors chose to adopt slightly different rhetorical strategies to effectively express their specific opinions in this specific communicative context. In my opinion, similar forms of fine-grained register variation generally account for the successful application of standard methods for stylometric authorship analysis, although additional research is needed to further evaluate this hypothesis, including analysing register variation in the writings of individual authors.

I have also argued that sociolinguistic explanations for the uniqueness of the language of individuals, as often proposed or implied in authorship analysis, are inconsistent with the basic methodological assumptions of stylometry. Stylometry is not based on the analysis of sociolinguistic variables, and therefore stylometry is not grounded in sociolinguistic theory, which in fact explicitly rejects the types of linguistic variables that are the focus of most research in stylometry. In the case study reported in this paper, the styles used by the authors under analysis were not distinguished with such a high degree of accuracy because they wrote in different dialects (i.e. because they used functionally equivalent variants of sociolinguistic variables at differential rates). These two authors, who come from very similar social backgrounds and who wrote for the same audience, were distinguished primarily because they chose to write opinion articles in slightly different registers, reflecting personal differences in the meanings they wished to express and how they wished to express those meanings.

In addition to considering how the results of this study, and the results of research in stylometry more generally, can be explained by linguistic theory, it is also important to consider how these results can inform our general theoretical understanding of the mechanisms of language variation and change. Although language is almost always embedded in society — a means for communication between individuals — its production is generally the act of a single individual. Ultimately language change must therefore be driven, in some way, by individuals changing language over time. A basic question is therefore why do individuals vary their language – a question that the study of individual variation in stylometry can certainly inform.

The standard view in variationist sociolinguistics is that language variation and change is driven by the expression of *social meaning*. Although there has been considerable debate about the nature of social meaning (Eckert 2012), it is clear that people's language varies depending on both their general social background and the specific social identities they choose to express through their use of language. Alternatively, the expression of *referential meaning* is generally seen as being independent of language change (Labov 2001). The underlying assumption is that the communicative function of language is stable over time and across languages and dialects. Change is seen as affecting the structure of language, allowing for different social meanings to be conveyed, but not as affecting the basic potential of language for communicating referential meaning. Variartionist sociolinguistics generally rejects functional explanations for language change.

This viewpoint, however, is difficult to reconcile with research on register variation, which has demonstrated that language has been adapted *over time* to maximise the communication of information across different communicative contexts (Biber and Conrad 2019; Grieve 2022). In particular, the multidimensional analysis of language change in specific communicative contexts has shown that the structure of language is actively adapted by people to allow for effective communication. For example, Biber and Gray (2016) show how the structure of academic writing has become more compressed over centuries so as to better allow for complex scientific information to be conveyed efficiently. Alternatively, Clarke and Grieve (2019) show how Donald Trump and his team adjusted their style of posting on Twitter over the course of the 2016 presidential election to adapt their language to the state of the campaign.

Although long ignored in sociolinguistics, research in stylometry provides another important perspective on this debate, offering direct empirical evidence that this type of linguistic adaptation is something that individuals do in variable ways. As this study has shown, people develop unique ways of using language – given the specific affordances and constraints of the situations in which they

interact and the specific referential meanings they wish to convey – so as to maximise their communicative efficacy in real communicative contexts, as Bloch implied when he introduced his more general conception of the idiolect.

That is not to deny the importance of the expression of social meaning: people also develop unique ways of using language to express the specific social meanings they wish to convey. Rather, it is to insist that the expression of referential meaning is an independent and important source of linguistic variation. When the study of linguistic variation is divorced from the study of the communication of referential meaning, as has been the methodological imperative of variationist sociolinguistics, the ability to observe relationships between communication and variation is naturally lost, but this does not mean that such relationships do not exist. As research on authorship and register variation has repeatedly shown, how people choose to use language to communicate meaning effectively is fundamental to the process of language variation and change.

Acknowledgements: I would like to thank Andrea Nini and David Wright for inviting me to present this paper at the *1st Roundtable on Practices and Standards in Forensic Authorship Analysis* at the University of Manchester in May 2019, as well as Jesse Egbert, Bethany Gray, and Tove Larsson for inviting me to submit this paper to a special issue in honour of Douglas Biber they are editing for *Corpus Linguistics and Linguistic Theory*. I also thank Tim Grant, Krzys Kredens, Andrea Nini, and Emily Waibel for discussing the contents of this paper with me, as well as the anonymous reviewers. Finally, I would like to especially thank Doug Biber for inspiring this work. Before coming to Northern Arizona University to study under his supervision in 2005, I had been working on authorship analysis. Doug introduced me to register analysis. In recognition of his deep influence on me as a linguist, this paper is my attempt, on the occasion of his retirement, to bring together these two distinct traditions of quantitative linguistic analysis, and to consider the wider implications of their unification on theories of language variation and change.

References

Argamon, Shlomo. 2018. Computational forensic authorship analysis: Promises and pitfalls. *Language and Law* 5(2). 7–37.

Baayen, Harald. 2001. *Word frequency distributions*. Dordrecht, Netherlands: Springer Science & Business Media.

Biber, Douglas. 1988. *Variation across Speech and writing*. Cambridge, UK: Cambridge University Press.

Biber, Douglas & Edward Finegan. 1994. Multi-dimensional analyses of authors' styles: Some case studies from the eighteenth century. In D. Ross & D. Brink (eds.), *Research in humanities computing*, vol. 3, 3–17. Oxford, UK: Oxford University Press.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge, UK: Cambridge University Press.

Biber, Douglas & Susan Conrad. 2019. *Register, genre, and style*. Cambridge, UK: Cambridge University Press.

Biber, Douglas & Bethany Gray. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes* 9(1). 2–20.

Biber, Douglas & Bethany Gray. 2016. *Grammatical Complexity in academic writing*. Cambridge, UK: Cambridge University Press.

Binongo, Jose. 2003. Who wrote the 15th book of oz? An application of multivariate analysis to authorship attribution. *Chance* 16(2). 9–17.

Bloch, Bernard. 1948. A set of postulates for phonemic analysis. *Language* 24(1). 3–46.

Bucholtz, Mary & Kira Hall. 2004. Language and identity. In Alessandro Duranti (ed.), *A Companion to linguistic anthropology*, 369–394. Malden, MA: Wiley.

Burrows, John. 2002. 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* 17(3). 267–287.

Clarke, Isobelle & Jack Grieve. 2019. Stylistic variation on the Donald Trump twitter account: A linguistic analysis of tweets posted between 2009 and 2018. *Plos One* 14(9). e0222062.

Coulthard, Malcolm. 2004. Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics* 25(4). 431–447.

Coulthard, Malcom, Alison Johnson & David Wright. 2016. *An Introduction to forensic linguistics: Language in evidence*. Abingdon, UK: Routledge.

Dauber v. Merrell Dow Pharmaceutical, Inc. 509 U.S. 579. 1993. 593–594.

Eckert, Penelope. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology* 41. 87–100.

Eder, Maciej, Rybicki Jan & Mike Kestemont. 2016. Stylometry with R: A package for computational text analysis. *R Journal* 8(1). 107–121.

Everitt, Brian & Torsten Hothorn. 2011. *An Introduction to applied multivariate Analysis with R*. Berlin, Germany: Springer Science & Business Media.

Grant, Tim. 2022. *The Idea of Progress in forensic authorship analysis*. Cambridge, UK: Cambridge University Press.

Grant, Tim & Kevin Baker. 2001. Identifying reliable, valid markers of authorship: A re- sponse to chaski. *Forensic Linguistics* 8(1). 66–79.

Grant, Tim & Nicci MacLeod. 2018. Resources and constraints in linguistic identity performance – a theory of authorship. *Language and Law* 5(1). 80–96.

Grant, Tim & Nicci MacLeod. 2020. *Language and online identities: The undercover Policing of internet sexual crime*. Cambridge, UK: Cambridge University Press.

Grieve, Jack. 2005. *Quantitative authorship attribution: A History and an Evaluation of techniques*. Burnaby, Canada: Simon Fraser University MA Dissertation.

Grieve, Jack. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing* 22(3). 251–270.

Grieve, Jack. 2016. *Regional variation in written American English*. Cambridge, UK: Cambridge University Press.

Grieve, Jack. 2022. Situational diversity and linguistic complexity. *Linguistic Vanguard*. https://doi.org/10.1515/lingvan-2021-0070.

Grieve, Jack, Tom Ruette, Dirk Speelman & Dirk Geeraerts. 2017. Social functional linguistic variation in conversational Dutch. In Eric Friginal (ed.), *Studies in corpus-based sociolinguistics*, 253–272. Abingdon, UK: Routledge.

Grieve, Jack & Helena Woodfield. 2021. Investigative linguistics. In Malcolm Coulthard, Alison May & Rui Sousa-Silva (eds.), *The Routledge handbook of forensic linguistics*, 2nd edn., 660–674. Abingdon, UK: Routledge.

Groscup, Jennifer L., Steven D. Penrod, Christina A. Studebaker, Matthew T. Huss & M. Kevin O'Neil. 2002. The effects of Daubert on the admissibility of expert testimony in state and federal criminal cases. *Psychology, Public Policy, and Law* 8(4). 339–372.

Halliday, M. A. K. 1978. *Language as social semiotic: The social Interpretation of Language and meaning*. London, UK: Edward Arnold.

Hockett, Charles F. 1958. *A Course in modern linguistics*. New York, USA: MacMillan Company.

Juola, Patrick. 2012. Stylometry and immigration: A case study. *Journal of Law and Policy* 21. 287–298.

Kestemont, Mike. 2014. Function words in authorship attribution from black magic to theory? In *Proceedings of the 3rd workshop on computational linguistics for literature*, 59–66.

Koppel, Moshe, Jonathan Schler & Shlomo Argamon. 2013. Authorship attribution: What's easy and what's hard? *Journal of Law and Policy* 21. 317–331.

Kredens, Krzysztof, Piotr Pezik & Lisa Rogers. 2019. Toward linguistic explanation of idiolectal variation – understanding the black box. In *Paper presented at the 14th biennial conference of the international association of forensic linguistics*, 1–5. Melbourne, Australia.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia, USA: University of Pennsylvania Press.

Labov, William. 2001. *Principles of language change: Internal factors*. Malden, MA: Wiley.

McMenamin, Gerald R. 2002. *Forensic linguistics: Advances in forensic stylistics*. Boca Raton, USA: CRC Press.

McMenamin, Gerald R. 2010. Forensic stylistics Theory and practice of forensic stylistics. In Malcolm Coulthard & Alison Johnson (eds.), *The routledge handbook of forensic linguistics*, 1st edn., 515–535. Abingdon, UK: Routledge.

Nini, Andrea. 2013. Codal variation theory as a forensic tool. In *Bridging the gap(s) between language and the law: Proceedings of 3rd European conference of the international association of forensic linguistics*, 31–41. Faculdade de Letras da Universidade do Porto.

Nini, Andrea. 2019. The multi-dimensional analysis tagger. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-dimensional analysis, 25 years on: A tribute to douglas biber*, 67–94. Amsterdam, Netherlands: John Benjamins.

Nini, Andrea. 2023. *A Theory of linguistic individuality for authorship identification*. Cambridge, UK: Cambridge University Press.

Nini, Andrea & Tim Grant. 2013. Bridging the gap between stylistic and cognitive approaches to authorship analysis using systemic functional linguistics and multidimensional analysis. *International Journal of Speech Language and the Law* 20(2). 173–202.

Pijpops, Dirk. 2020. What is an alternation?: Six answers. *Belgian Journal of Linguistics* 34(1). 283–294.

Preston, Dennis. 2001. Style and the psycholinguistics of sociolinguistics: The logical problem of language variation. In Penelope Eckert & John Rickford (eds.), *Style and sociolinguistic variation*, 279–304. Cambridge, UK: Cambridge University Press.

Sardinha, Tony Berber & Marcia Veirano Pinto (eds.). 2014. *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber*. Amsterdam, Netherlands: John Benjamins.

Sardinha, Tony Berber & Marcia Veirano Pinto (eds.). 2019. *Multi-dimensional analysis: Research methods and current issues*. London, UK: Bloomsbury.

Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60(3). 538–556.

Tagliamonte, Sali. 2011. *Variationist sociolinguistics: Change, observation, interpretation*. Malden, MA: Wiley Blackwell.

Taylor, Gary, John Jowett, Terri Bourus & Gabriel Egan (eds.). 2016. *The new Oxford Shakespeare: Modern critical edition*. Oxford, UK: Oxford University Press.

Taylor, Gary & Gabriel Egan (eds.). 2017. *The new Oxford Shakespeare: Authorship companion*. Oxford, UK: Oxford University Press.

Wright, David. 2017. Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. *International Journal of Corpus Linguistics* 22(2). 212–241.