Vol. 4, No. 1, 2019

Journal of
language | works

"Because the computer said so!" Can computational authorship analysis be trusted?

Alesia Locker
Aarhus University
201603388@post.au.dk

Abstract

This study belongs to the domain of authorship analysis (AA), a discipline under the umbrella of forensic linguistics in which writing style is analysed as a means of authorship identification.

Due to advances in natural language processing and machine learning in recent years, interest in computational methods of AA is gaining over traditional stylistic analysis by human experts. It may only be a matter of time before the software will assist, if not replace, a forensic examiner. But can we trust its verdict? The existing computational methods of AA receive critique for the lack of theoretical motivation, black box methodologies and controversial results, and ultimately, many argue that these are unable to deliver viable forensic evidence.

The study replicates a popular algorithm of computational AA in order to open one of the existing black boxes. It takes a closer look at the so-called "bag-of-words" (BoW) approach – a word distributions method used in the majority of AA models, evaluates the parameters that the algorithm bases its conclusions on and offers detailed linguistic explanations for the statistical results these discriminators produce.

The framework behind the design of this study draws on multidimensional analysis – a multivariate analytical approach to linguistic variation. By building on the theory of systemic functional linguistics and variationist sociolinguistics, the study takes steps toward solving the existing problem of the theoretical validity of computational AA.

Keywords: forensic linguistics; authorship; variation; NLP.

1. Introduction

We live today in a fast-paced information age, where any person can communicate instantly and publish any text to a potentially global audience through the interwoven networks of blogs, social media and instant news platforms. In this online reality, individuals are free to keep their true identity private, to actively conceal themselves, or to appear under alternative internet personas. On this new front of human activities, there is a growing need for law enforcement and intelligence agencies to respond to individuals and organised groups who exercise this anonymity with malicious and manipulative intentions or with an intent to impersonate others.

From phishing email scams to international disinformation campaigns by foreign powers, such activities have become increasingly sophisticated and untraceable, where the text itself is often the only clue to its author. In light of this issue, the development of intelligent, efficient and reliable methods of authorship analysis is increasingly becoming a question of practical interest.

The term authorship analysis (AA) refers to one of the routine tasks in the domain of forensic linguistics (Coulthard et al., 2016), an area of applied linguistics concerned with the use of language in a legal context and the provision of linguistic evidence. AA can be defined as "the task of inferring characteristics of a document's author, including but not limited to identity, from the textual characteristics of the document itself" (Juola, 2007, p. 120). In this respect, the text becomes a piece of evidence in a legal case.

AA has a long history, from human-driven decisions to algorithmically-driven ones. To date, the traditional stylistic approach to AA is the one most often employed in forensic casework, due to its explanatory potential and its explicit correlation within the well-established forensic area of handwriting examination.

For example, in the case of the *murder of Amanda Birks,* Amanda's husband, Christopher, attempted to disguise the actual cause of his wife's death as an accident by sending a number of text messages from her phone. The analysis of spelling variants allowed the forensic examiner to conclude that the text messages were forged and pointed toward Christopher as the actual author (Grant, 2013). Christopher later confessed and was charged with murder. In another similar homicide case, a husband forged his wife's farewell letter on a typewriter but was identified through spelling errors (Eagleson, 1994).

The methods of the stylistic framework are not clearly defined since prior to the examination of the text, examiners can never precisely know what sort of features they are looking for (Grant, 2013, p. 472). In his book on forensic stylistics, McMenamin (2002) presents an extensive list of potential style markers based on his experience with 80 authorship cases. The markers include, but are not limited to text format, usage of numbers and symbols, abbreviations, punctuation, capitalisation, spelling, word formation, syntax, discourse, errors and corrections, and high-frequency words.

Examiners typically rely on a combination of qualitative and quantitative techniques to identify the recurring patterns across the compared texts and to ensure the highest recovery of any forensically relevant information. However, due to the lack of established protocol, potential subjectivity, and a risk of confirmation bias (Nickerson, 1998; Risinger et al., 2002), AA is not always admissible in court testimonies (Coulthard, 2013). Such analysis is also very time- and resource-consuming and therefore inadequate for problems involving greater amounts of textual data within limited timeframes, as is common in contemporary scenarios.

Computational AA, or stylometry, on the other hand, seems to be a solution to these problems. These days, most computational linguists have experimented with creating a model of some sort that is able to classify texts by author. The result is a numerous array of available methods and programmes. According to one of these open sources, "one of the primary uses of the stylometric software is verifying the authorship of a text, whether for academic purposes or to expose a forgery" (Emma Identity, 2017).

Attempts have been made to take AA technology to court (Chaski, 2013), albeit not without critique, and there are ongoing attempts to implement it in security settings (see the research by Privacy, Security and Automation Lab). But the question is whether we can trust its verdict.

2. Stylometric framework of authorship analysis

Stylometric methods of AA are a group of computational approaches to the analysis of authorship, originally from the domain of digital humanities. Proponents of stylometry attempt to extract forensically viable stylistic information from the frequent linguistic patterns recurring in texts by the same author (Stamatatos, 2013, p. 423).

The methods were originally developed for the analysis of disputed authorship in historical literary studies. As far back as in 1887, Mendenhall suggested attributing authorship based on word length distributions. The method required calculating the frequency distributions of the words of various lengths in a text (number of one-grapheme words, two-grapheme words, three-grapheme words, etc.) and comparing the pattern that emerged to other samples of writing. In a later study, Mendenhall (1901) made an attempt to shed light on the authorship of Shakespeare plays by comparing texts by Shakespeare, Marlowe and Bacon using the same method. Although word length did not prove to be a reliable indicator of authorship, being highly susceptible to the register and subject, the methodological novelty and the scope of work conducted in the course of the study deserves credit: in those days, all the calculations were done by hand.

Contemporary stylometry consists primarily of machine learning techniques employing statistical pattern recognition – the degree of similarity between texts is measured by mathematical models on the basis of specific sets of frequently recurring features (see reviews by Juola, 2008; Koppel et al., 2009; Rocha et al., 2017; Stamatatos, 2009).

In the process of stylometric AA, the texts are broken down into their components (character and word level sequencies of *n* elements – n-grams) which are then quantified, and their frequencies further serve as stylistic discriminators. The typical features that the models rely on to match "stylistic fingerprints" (Stamatatos et al., 2016) are lexical components and measurements of text complexity such as vocabulary richness, and average word and sentence length.

Tracing authorship based on distributional patterns of lexical features gained its popularity due to them being the easiest ones to extract and, at the same time, facilitating high accuracies in AA models. The texts are either split into individual words or graphemes (e.g. Kocher & Savoy, 2017) or their sequences of a certain length – n-grams (e.g. Koppel & Winter, 2014). Although contemporary automated grammatical and syntactic text markup (tagging and parsing) allow extraction and quantification of information on the use of syntactic structures and grammatical categories, it has been empirically established that a simple bag-of-words (BoW) approach (frequent words of the dataset) and character n-grams (sequences of *n*-graphemes) are the most accurate stylistic discriminators to date (surveys by Grieve, 2007; Stamatatos et al., 2016).

The process of stylometric AA can be briefly summarised as follows:

1. The texts to be analysed are pre-processed to make them readable by software. This stage can also include morphosyntactic tagging.
2. An algorithm extracts a certain set of pervasive linguistic features (e.g. character or word n-grams, punctuation characters, parts of speech), calculates the frequency of each feature and represents each text as a matrix of these standardised frequencies.
3. This matrix is further subjected to an unsupervised or a supervised machine learning algorithm which examines the similarities and differences between the individual texts and further groups statistically similar texts together.

The approach has received a lot of attention in recent years due to PAN @ CLEF evaluations – a conference on computational AA where attendance is possible after successfully completing several AA tasks. The contemporary models have been successfully applied to large data samples, and there is active research going on to expand their use to shorter texts and colloquial registers (see, e.g., Abbasi & Chen, 2005; Rocha et al., 2017; Zheng et al., 2006).

Contemporary algorithms have dramatically improved from the first attempts at capturing the idiolectal style through quantification and are now argued to consistently achieve high accuracy over a variety of AA problems. The indisputable advantage of computational analysis is its objectivity, that is to say that conclusions are not being based on the subjective judgement of an analyst. Potentially, any stylometric method is replicable, its error rate is tested, and the weight of the evidence (authorship likelihood) may be estimated statistically.

However, a pitfall of stylometric methods is that there is no linguistic theory that supports the validity of this framework. From a linguistic point of view, for example, it is unclear why combinations of alphabetic characters or most frequent words of a text should represent an idiolectal (individual) writing style. Essentially, stylometry fails to provide an explanation at the core level: why is it that the reference set of features used in the model should be able to discriminate idiolectal styles?

As a deductive analytical approach, stylometry is primarily concerned with testing hypotheses on whether a particular category of quantifiable linguistic variables could differentiate between the authors, rather than explaining why they should be reliable indicators of authorship. If a particular feature set has consistently achieved high prediction accuracy over various authors, they are considered reliable stylistic discriminators for further application to real-world AA problems. From the stylometry sceptics' point of view, this is seen as a way of dropping "the difficult task of identifying reliable and valid markers of authorship" and side-stepping it "by simultaneously testing a range of statistics to identify discriminants that work in a particular case" (Grant & Baker, 2001, p. 76).

It is well known that while statistics may give a sense of objectivity, it does not always produce results that can be trusted. What is worse, stylometric models are essentially "black boxes" as far as the user is concerned, and do not allow for evaluating linguistic evidence per se (Cheng, 2013, p. 547; Solan, 2013, p. 567). Basically, we put in disputed and undisputed samples of writing and then get the result of authorship attribution as an output. What happens in between these two stages is concealed from outside scrutiny. Due to the complexity of the decision-taking algorithm, it is not really possible to inspect why it reached the conclusions it reached. We suppose that they are accurate ones, based on previous tests. Obviously, this is not a satisfactorily rigorous argument in criminal justice settings.

Rather than coming up with a new algorithm for AA, this study will attempt to open up one of the existing black boxes. By conducting an experiment of our own, in the following sections we will look at what constitutes the common BoW method and whether using word distributions as indicators of authorship can be justified.

3. Experimental setup

The BoW model is commonly used in text classification tasks. Within this computational model, each text is represented as a set of lexical items disregarding the context of their use. The frequency of each word is computed and used as a feature for training a text classifier (McTear et al., 2016, p. 167).

The number of the words used in stylometric models is arbitrary, with the research suggesting various strata ranging from 50 most frequent words (MFW) to all the words in the data set. A detailed survey of the method was done by Kestemont (2014) in the article "Function Words in Authorship Attribution. From Black Magic to Theory?", although the article actually does not explicitly name any linguistic theory which would support its premise.

The experiment in this paper follows a typical stylometric pipeline. All the computations and feature extraction were carried out in R (R Core Team, 2017), a programming language for statistical computing, implementing *Stylo* package (Eder et al., 2016).

3.1 Data set

The data set for this study was compiled in such a way as to ensure that no significant differences between the texts could be attributed to factors other than a variation of authors.

The two authors chosen for the study are columnists of *The Guardian*, a British daily newspaper. The texts of this newspaper are routinely used for the task of author attribution and clustering in PAN @ CLEF evaluations, a competition on authorship attribution, for a number of reasons. The data set satisfies the criterion of ground-truth data that is necessary to assess the accuracy of the AA method (Chaski, 2013, p. 336) since the authors are verifiable, and the texts are authentic samples of writing free from the observer's paradox (Labov, 1972, p. 209). By reason of the data coming from the same media source, external editing supposedly affects all texts to a similar extent (if at all) and will not be a crucial factor responsible for the stylistic differences among the texts (Stamatatos, 2013, p. 430).

Andrew Rawnsley and Matthew d'Ancona publish opinion pieces in similar thematic areas (UK politics, Table 1) and have similar sociolinguistic backgrounds. Both are male, are of British origin and currently live in London. Presumably, if the analysis could distinguish between the writings of this pair of authors, it should be successful in other cases.

For the purpose of this experiment, a total of 20 texts by each author were extracted. The length of the texts in the data set ranges from 796 to 1618 words.

Table 1. Distribution of topics in the data set

Topic	Rawnsley	d'Ancona	Total
Theresa May	4	2	6
Brexit	4	4	8
Politics	3	2	5
Conservatives	1	1	2
Budget 2017	1	1	2
Jeremy Corbyn		1	1
Economic Policy	1		1
Catalonia		1	1
Aid		1	1
Trump Administration		1	1
Freedom of Speech		1	1
Donald Trump	1	1	2
The Handmaid's Tale		1	1
Domian Green		1	1
Counter-terrorism Policy		1	1
Conservative Conference		1	1
Sexual Harassment	1		1
Productivity	1		1
Oxfam	1		1
Momentum	1		1
Housing	1		1
Total	*20*	*20*	*40*

3.2 Data annotation

The corpus was annotated using the Multidimensional Analysis Tagger (MAT) (Nini, 2014) for 67 lexicogrammatical features, which proved useful for the investigation of variation in English language in a multidimensional framework by Biber (1995, p. 94), grounded in systemic functional grammar (Halliday, 1985). The categories include, but are not limited to, tense and aspect markers, modals, nominal forms, adverbials, subordination features and lexical classes (Table 3).

Several other tags were manually added to refine *that*-categories. Adding the tags was deemed important due to the categories below often being used as stance devices (Biber et al., 2015, p. 313) as well as indicating marked word order choices:

68. *That* subject complement Take the fact that she gave a speech at all.
69. *That* clause of degree: The word *nice* has had such a bad press that my fingers hesitate to type it.
70. Subject predicative *that*-clause: The answer is that he was once a success.

3.3 Feature extraction

The number of occurrences of each of the words in a data set was then calculated and, following other stylometric studies, 100 words of the highest frequency were selected. The contractions (such as *it's* instead of *it is*) were preserved in order to retain the variability of use.

The wordlist was then reviewed to exclude topic-related content words, which appear in the list with a decline in frequencies in order to prevent categorisation of texts by the topic rather than the authorial style. This resulted in third-person pronouns (*he*, *his*, *she*, *her*, *they*, *their*), topic-related nouns (*Brexit*, *labour*, *May*, *party*, *Tory*, *EU*, *government*, *Britain*, *election*, *MPs*, *minister*, *Trump*, *Tories*, *voters*, *leader*) and adjectives (*political*, *prime*, *British*) being exempt from the analysis.

Although in stylometry all personal pronouns are customarily eliminated from the analysis as heavily content-dependent features (see Kestemont, 2014, p. 61), in this study the pronoun *it* and first- and second-person pronouns were retained as potentially useful: the singular pronoun *I* is commonly used in association with the expression of author stance, while the others frequently function as a generic reference rather than making actual referential links (Biber et al., 2015, pp. 94–96).

The analysis also includes both upper- and lower-case variants which would allow tracing preferences for the sentence-initial position (e.g. sentence-initial coordinators *and*, *but* are generally dispreferred in written registers (Biber et al., 2015, p. 229)). This resulted in the list being doubled to 200. The list was then culled at 50 % to preserve only those words that occurred in at least half of the texts of the corpus, to avoid sparse data. As a result, a matrix of the data set containing the relative frequencies of 106 lexical variables across 40 texts was obtained. As shown in Table 2, the majority of the lexical variables retained in the analysis are function words.

Table 2. MFW: lexical features considered in authorship analysis

The	it	has	so	all	because	last	being
the	for	are	its	we	into	made	right
of	As	have	had	can	like	any	did
to	as	an	been	some	much	say	does
A	be	at	when	those	do	after	make
a	on	from	I	now	many	too	own
And	was	will	If	most	our	way	before
and	with	would	if	them	such	should	never
In	by	who	what	you	just	might	
in	But	more	than	out	other	against	
That	but	There	which	up	over	rather	
that	not	there	no	only	how	still	
is	This	or	were	could	then	these	
It	this	one	about	even	also	where	

The relative frequencies of the functional categories were obtained as an output of *MAT* Analyser (Nini, 2014). Again, only the features which occurred in at least 50 % of the texts of the corpus were retained, which resulted in 61 most frequent tags (MFT) of the data set being left in the list (Table 3).

Table 3. MFT: functional categories considered in authorship analysis

Tag	**Description**	**Tag**	**Description**
VBD	past tense	PIRE	pied-piping relative clause
PEAS	perfect aspect	CAUS	causal adverbial subordinator
VPRT	present tense	COND	conditional adverbial subordinator
PLACE	place adverbials	OSUB	other adverbial subordinator
TIME	time adverbials	PIN	total prepositional phrases
FPP1	first-person pronoun	JJ	attributive adjectives
SPP2	second-person pronoun	PRED	predicative adjectives
TPP3	third-person pronoun	RB	total adverbs
PIT	pronoun *it*	WHSUB	*Wh* relative clause on subject position
DEMP	demonstrative pronoun	THATD	subordinator *that* deletion
INPR	indefinite pronoun	CONJ	conjuncts
PROD	pro-verb *do*	DWNT	downtoners
WHQU	*Wh* questions	AMP	amplifiers
NOMZ	nominalisation	EMPH	emphatics
GER	gerund	DEMO	demonstratives
NN	total other nouns	POMD	possibility modals
PASS	agentless passive	NEMD	necessity modals
BYPA	*by*-passive	PRMD	predictive modals
BEMA	*be* as main verb	PUBV	public verbs
EX	existential *there*	PRIV	private verbs
THVC	*that* verb complements	SUAV	suasive verbs
THAC	*that* adjective complement	SMP	*seem* and *appear*
THSC	*that* subject complement	CONT	contractions

TSPRED	*that* subject predicative clause	STPR	stranded preposition
WHCL	*Wh* clause	SPAU	split auxiliaries
PRESP	present participial adverbial clause	PHC	*and* phrasal coordination
PASTP	past participial adverbial clause	ANDC	*and* independent clause coordination
WZPAST	past participial postnominal clause	CC	other coordinators *(but, or)*
WZPRES	present participial postnominal clause	SYNE	synthetic negation
TSUB	*that* relative clause on subject position	XX0	analytic negation
TOBJ	*that* relative clause on object position		

3.4 Principal Component Analysis: a case study

To conduct the authorship analysis and compare the styles of the authors pairwise, the experiment utilises an exploratory statistical technique – principal component analysis (PCA).

PCA is a powerful multivariate statistical tool for dimensionality reduction. Essentially, it allows us to visualise relatedness between the set of observations (texts characterised by multiple parameters) through a simplified two-dimensional representation of their relative positions by plotting them on a single scatter plot.

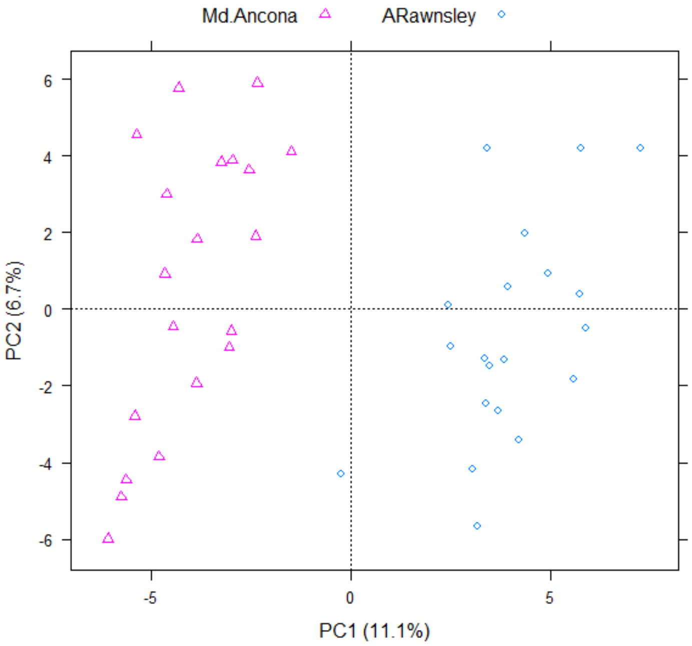

Figure 1. PCA: d’Ancona vs Rawnsley, 167 lexicogrammatical variables

Figure 1 visualises stylistic distance and relatedness between writings by d’Ancona and Rawnsley. The first principal component (PC1), which accounts for the most variance in the data, reveals a clear separation of the texts by author, suggesting that the linguistic variables considered in this study indeed capture idiosyncratic language use. Out of the total of 40 texts, there is one outsider, a text on politics, which the analysis fails to attribute to Rawnsley’s grouping, indicating that it lacks some of his most prominent stylistic features. Neither does it possess d’Ancona’s style markers, and therefore it does not get erroneously clustered with his pieces of writing.

The topic of the texts, which is generally one of the major factors determining vocabulary choices, does not seem to have a significant impact on the clustering process. The text groupings in Figure 1 reflect inter-author variation rather than similarities and differences in topic between the texts, and therefore support the argument that the MFW used in this analysis are relatively topic independent.

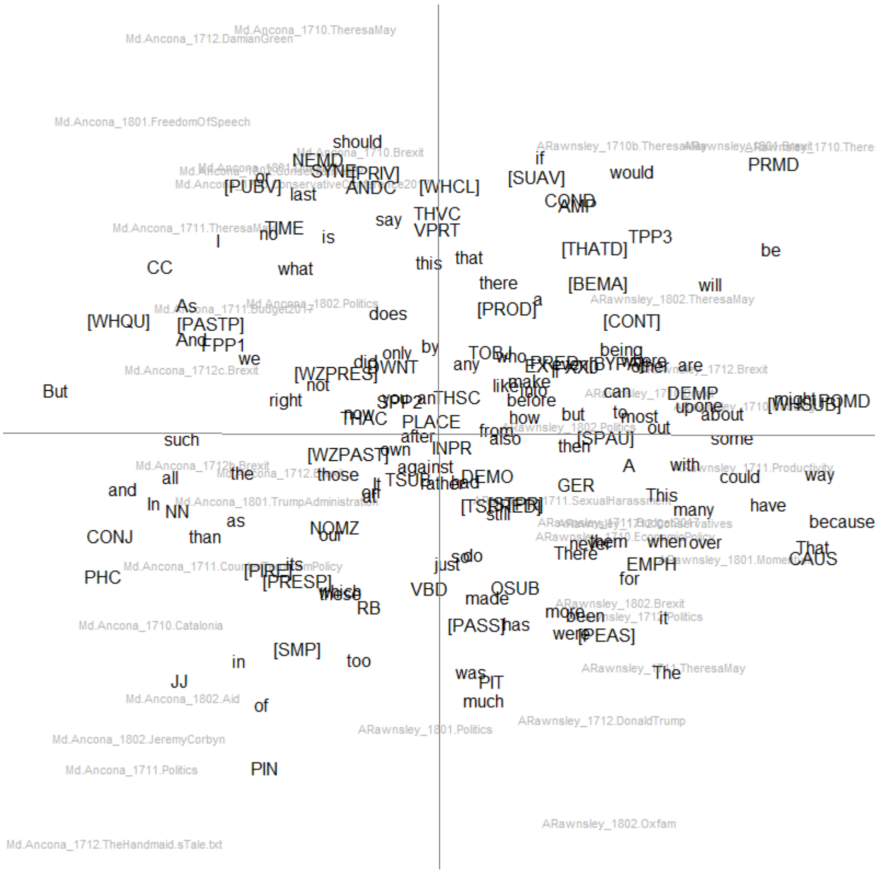

Figure 2. PCA: contrasting use of the 167 lexicogrammatical variables by d’Ancona and Rawnsley.

Figure 2 represents the loadings of the principal components – a correlation between the texts and the 167 lexicogrammatical variables considered in the analysis, revealing the contrasting use of these linguistic features by d’Ancona and Rawnsley. The texts with the high values for PC1 (by Rawnsley) would be expected to have high values for the variables at the right of the spectrum of PC1, whereas the texts with low values on PC1 (by d’Ancona) would have low scores for each of those variables and high scores for the variables at the left of the spectrum.

Examination of the component loadings reveals an interesting observation. The MFW used in this experiment can be seen as lexical variants of Biber’s functional grammatical categories, which are commonly used by proponents of the multidimensional framework (Biber 1988) to differentiate between registers and styles in functional terms. The grouping of grammatical and lexical variables on Rawnsley’s end of PC1 (modals and causal conjunctions) is linked to sharing personal attitudes while conveying information. D’Ancona’s style, on the contrary, is characterised by the frequent occurrence of features associated with informational production (Biber, 1995, p. 142): nouns, attributive adjectives and phrasal coordination.

Modals *(could*, *will*, *would)* and the causal adverbial subordinator *because*, which frequently occur in the texts by Rawnsley, are characterised by a shared function – creating frames for argumentation. The modals express the author's stance towards the propositions, while the subordinator *because* marks causal relationship. To a large extent, the use of the word *way*, which occurs in all the texts by Rawnsley (39 instances) and only eight times (in 8 texts) in d'Ancona's data, reflects this practice. By presupposing the existence of alternatives, the word takes its place in the overall scheme of Rawnsley's argumentation and evaluation. Below are 4 of 39 concordance lines for the word *way* in Rawnsley's data:

(11)

N	Concordance	
1	to punish that debacle in the traditional	way ***because*** they feared that bad could be followed
2	strike for the throne. That leaves one other	way out of this grisly stalemate, which ***would*** be
		(Rawnsley_1710-TheresaMay)
3	believing that it ***could*** be managed in a	way that contained the damage to trade, jobs and
4	is essentially futile to try to find a	way forward ***because*** Mrs May isn't capable of
		(Rawnsley_1710-Brexit)

Both authors use sentence-initial *That* as a determiner and demonstrative pronoun to express their stance and attitudes towards propositions, although with significant variation in frequencies: *That* occurs 53 times in Rawnsley's data (19 texts) and only seven times (4 texts) in the data of d'Ancona. Consider the authors' concordances of *That* from the articles on the same topic:

(2)

N	Concordance	
1	believe that - but it would be manageable.	That view was especially prevalent in the upper
2	had no, or little, idea who they were.	That makes it even less excusable. Mrs May was
3	and contribute so much to society".	That's a sweet idea, but I can't
4	of the world from our connections to both.	That intricate and subtle matrix, built up over
		(Rawnsley_1712-DonaldTrump)

(3)

N	Concordance	
1	his charm when handling his subjects.	That's precisely why no remotely competent adviser
		(d'Ancona_1801-DonaldTrump)

Still, there is a stylistic device associated with argumentation and persuasion which d'Ancona tends to use more frequently than Rawnsley – rhetorical questions. Although both authors use rhetorical questions to emphasize the point being discussed (examples 4 and 5), these questions occur in 17 out of 20 d'Ancona's articles and in 7 out of 20 articles by Rawnsley:

(4)

> But there is a greater and more pernicious force at work here, too - and one that is not confined to the Labour party. It is the notion that we live in an age of daily reboot, in which the past is of only antiquarian interest and certainly of little relevance to the present. **The cold war? Didn't that all end in 1989? Why drag that up?**
>
> To which the answer is - because the jagged legacy of that exhausting struggle does much to explain the predicament in which we find ourselves today.
>
> (d'Ancona_1802-JeremyCorbyn)

(5)

> This has put us in the highly unusual position where neither the Conservative frontbench nor its opposition counterpart is leading the national debate. **How is it that Anna Soubry, an ex-minister, and Jacob Rees-Mogg, a never-was minister, have become so prominent in the arguments?** It is at least in part because they are filling the vacuum left by a government strangled by its internal divisions and an official opposition made inarticulate by its unwillingness to take robust positions.
>
> (Rawnsley_1802-Politics)

Example (4) demonstrates another prominent trace of d'Ancona's authorship – a sentence-initial coordinator *But*. Generally, in both authors' data, the coordinator *but* is characterised by very similar distributions. Yet, Rawnsley begins a sentence with *But* 12 times out of 120 instances of his use of this coordinator. D'Ancona, on the contrary, has a stronger preference for the sentence-initial variant. In his data, the coordinator occupies the sentence-initial position in 73 out of its 129 occurrences.

The frequencies of the individual features, even the most pervasive ones, may not be completely reliable indicators of authorship as such (e.g., the abovementioned sentence-initial coordinator *But* whose occurrence in a text does not necessarily indicate that it belongs to d'Ancona). But since the multivariate statistical analysis considers co-occurring patterns, this significantly informs the model, allowing it to draw a rather sharp separation between the two styles.

4. Discussion

It has been empirically established in stylometry, and once again supported by this experiment, that the most frequent words of the data set can be useful in tracing authorship.

Stylometry seems to have arrived at the same findings as a discipline from the domain of corpus linguistics – multidimensional analysis, which argues that any register or style can be distinguished by the most pervasive functional features, present in all texts but characterised by very specific distributions in each category (Biber & Conrad, 2009, pp. 55–56).

Studies into linguistic variation utilising the multidimensional framework developed by Douglas Biber (1988) have demonstrated that dialects and jargons of social groups can be distinguished by very basic quantifiable lexicogrammatical patterns – the frequencies of occurrence of functional forms of the language and their co-occurrence. Although function word classes and grammatical constructions of a language in principle are available to all its users, their distributions often vary significantly in the speech and writing of individuals and social groups (Biber & Finegan, 1994; Biber & Conrad, 2001). The preferences of language users for certain lexicogrammatical variables correlate significantly with social dimensions such as age, gender, ethnicity, education, socioeconomic class, social network, etc. (Hancock et al., 2013; Helt, 2001; Milroy & Milroy, 1998; Pennebaker, 2011).

It has been established (Tagliamonte, 2012, p. 45) that individuals acquire certain frequencies of linguistic variables from their caregivers. From childhood through adolescence these frequencies undergo fluctuations, gradually aligning with the norms of the linguistic communities that the individuals are exposed to (Baxter & Croft, 2016) which results in language variation at the individual level. Since the combined set of factors that determine one's socio-linguistic background is unique to each language user, it can be expected to result in an idiosyncratic formula of stylistic preferences, making individual writing styles distinct and identifiable.

It seems that variationist sociolinguistics and systemic functional grammar could help us to gain an insight into the nature of the stylistic distinction between individuals and provide a relevant linguistic framework for computational AA. The multidimensional framework, grounded in systemic functional grammar, provides a missing theoretical motivation behind the successful performance of the BoW approach, suggesting that pervasive linguistic variables capture idiosyncratic language use (Biber & Conrad, 2009, pp. 71–72). The most frequent topic-independent words of the corpus routinely used in a BoW approach in stylometry are lexical variants of functional grammatical categories, closely linked to the authors' linguistic habits and self-expression.

5. Conclusion

In his article and critique of stylometry, Rudman (2012) argues that the setbacks and risks of computational AA may outweigh its benefits. Exploring the state-of-the-art of AA, Juola (2008) takes a more optimistic approach, suggesting that computational AA is possible, given that the examiner is competent and aware of all of its possible limitations.

The findings of this study support the latter proposition. Statistical techniques and computational tools can be employed to support a forensic examiner in the search for stylistic links in an extensive collection of texts. There is still considerable scope for research concerning the use of lexicogrammar in computational AA.

As for the existing software, it has to be taken with caution – not only that it is error-free, but also that it requires a sound understanding of the principles of computational linguistics to be used accurately, and that the statistical results should then be assessed thoughtfully.

Bibliography

Abbasi, A., & Chen, H. (2005) Applying Authorship Analysis to Arabic Web Content. In P. Kantor, G. Muresan, F. Roberts, D. D. Zeng, F. Wang, H. Chen & R. C. Merkle (Eds), *Intelligence and Security Informatics. ISI 2005* (pp. 183-197). Berlin/Heidelberg: Springer.

Baxter, G., & Croft, W. (2016). Modeling language change across the lifespan: Individual trajectories in community change. *Language Variation and Change, 28*(2), 129-173. doi:10.1017/S0954394516000077

Biber, D. (1988). *Variation across speech and writing* (Paperback ed.). Cambridge: Cambridge University Press.

Biber, D. (1995). *Dimensions of register variation: a cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, D., & Conrad, S. (Eds.). (2001). *Variation in English: Multi-Dimensional Studies*. Harlow: Pearson Education/Longman.

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Leech, G. N. (2015). *Longman student grammar of spoken and written English*. Harlow, Essex: Longman.

Biber, D., & Finegan, E. (Eds.). (1994). *Sociolinguistic Perspectives on Register*. New York: Oxford University Press.

Chaski, C. E. (2013). Best practices and admissibility of forensic author identification. *Journal of Law and Policy*, *21*(2), 333–376.

Cheng, E. K. (2013). Being pragmatic about forensic linguistics. *Journal of Law and Policy*, *21*(2), 541–550.

Coulthard, M., Johnson, A., & Wright, D. (2016). *An Introduction to Forensic Linguistics: Language in Evidence*. London/New York: Taylor & Francis.

Coulthard, M. (2013). A failed appeal. *International Journal of Speech Language and the Law*, *4*(2), 287–302. https://doi.org/10.1558/ijsll.v4i2.287

Eagleson, R. (1994). Forensic analysis of personal written texts: a case study. In J. Gibbons (Ed.), *Language and the Law* (pp. 362–373). Harlow: Longman.

Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *R Journal*, *8*(1), 107–121.

Emma Identity. (2017). Get Inside My Time Machine: A Quick Trip to the Stylometry Origin. Retrieved January 10, 2019, from https://medium.com/emma-identity/get-inside-my-time-machine-a-quick-trip-to-the-stylometry-origin-b65481549096

Grant, T. (2013). TXT 4N6: method, consistency, and distinctiveness in the analysis of sms text messages. *Journal of Law and Policy*, *21*(2), 467–494.

Grant, T., & Baker, K. (2001). Identifying reliable, valid markers of authorship: a response to Chaski. *International Journal of Speech, Language and the Law*, *8*(1), 66–79.

Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, *22*(3), 251–270. https://doi.org/10.1093/llc/fqm020

Halliday, M. A. K. (1985). *An introduction to functional grammar*. London: Edward Arnold.

Hancock, J. T., Woodworth, M. T., & Porter, S. (2013). Hungry like the wolf: A word-pattern analysis of the language of psychopaths. *Legal and Criminological Psychology*, *18*(1), 102–114. https://doi.org/10.1111/j.2044-8333.2011.02025.x

Helt, M. E. (2001). A multi-dimensional comparison of British and American spoken English. In D. Biber & S. Conrad (Eds.), *Variation in English: multi-dimensional studies* (pp. 171–183). Harlow: Pearson Education/Longman.

Juola, P. (2007). Future Trends in Authorship Attribution. In P. Craiger & S. Shenoi (Eds.), *Advances in Digital Forensics III* (pp. 119–132). Springer New York.

Juola, P. (2008). *Authorship Attribution*. Hanover, MA, USA: Now Publishers Inc.

Kestemont, M. (2014). Function Words in Authorship Attribution. From Black Magic to Theory? In A. Feldman, A. Kazantseva & S. Szpakowics (Eds.) *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)* (pp. 59–66). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-0908

Kocher, M., & Savoy, J. (2017). A simple and efficient algorithm for authorship verification. *Journal of the Association for Information Science and Technology*, *68*(1), 259–269. https://doi.org/10.1002/asi.23648

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Computational Methods in Authorship Attribution*, *60*(1), 9–26.

Koppel, M., & Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, *65*(1), 178–187. https://doi.org/10.1002/asi.22954

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania.

McMenamin, G. R. (2002). *Forensic linguistics: Advances in forensic stylistics*. Boca Raton, Fla: CRC Press.

McTear, M., Callejas, Z., & Griol, D. (2016). *The Conversational Interface: Talking to Smart Devices* (1st ed.). Springer Publishing Company, Incorporated.

Mendenhall, T. C. (1887). The Characteristic Curves of Composition. *Science*, *ns-9*(214S), 237. https://doi.org/10.1126/science.ns-9.214S.237

Mendenhall, T. C. (1901). A Mechanical Solution of a Literary Problem. *The Popular Science Monthly*, *LX (7)*, 97–105.

Milroy, J., & Milroy, L. (1998). Varieties and Variation. In F. Coulmas (Ed.), *The Handbook of Sociolinguistics*. Blackwell Publishing. Retrieved from <http://www.blackwellreference.com/subscriber/tocnode.html?id=g9780631211938_chunk_g97806312119385>

Nickerson, R. S. (1998). Confirmation Bias : A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Nini, A. (2014). *Multidimensional Analysis Tagger 1.3 - Manual*. Retrieved from http://sites.google.com/site/multidimensionaltagger

Pennebaker, J. (2011). *The secret life of pronouns: What our words say about us*. New York: Bloomsbury Press.

Privacy, Security and Automation Lab. (n.d.). Retrieved January 4, 2019, from https://psal.cs.drexel.edu/index.php/Main_Page

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Risinger, D. M., Saks, M. J., Thompson, W. C., & Rosenthal, R. (2002). The Daubert/Kumho Implications of Observer Effects in Forensic Science: Hidden Problems of Expectation and Suggestion. *California Law Review*, *90*(1), 1–56.

Rocha, A., W. J. Scheirer, C. W. Forstall, T. Cavalcante, A. Theophilo, B. Shen, E. Stamatatos. (2017). Authorship Attribution for Social Media Forensics. *IEEE Transactions on Information Forensics and Security*, *12*(1), 5–33. https://doi.org/10.1109/TIFS.2016.2603960

Rudman, J. (2012). The State of Non-Traditional Authorship Attribution Studies-2012: Some Problems and Solutions. *English Studies*, *93*(3), 259–274. https://doi.org/10.1080/0013838X.2012.668785

Solan, L. M. (2013). Intuition versus algorithm: the case of forensic authorship attribution. *Journal of Law and Policy*, *21*(2), 551–576.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.*, *60*(3), 538–556.

Stamatatos, E. (2013). On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, *21*(2), 421–439.

Stamatatos, E., Tschuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., & Potthast, M. (2016). Clustering by Authorship Within and Across Documents. In K. Balog, L. Cappellato, N. Ferro & C. Macdonald (Eds.), *CLEF*. Retrieved from http://ceur-ws.org/Vol-1609/16090691.pdf

Tagliamonte, S. (2012). *Variationist sociolinguistics: change, observation, interpretation*. Malden, MA: Wiley-Blackwell.

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, *57*(3), 378–393. https://doi.org/10.1002/asi.20316