The Stylometric Processing of Sensory Open Source Data

A thesis submitted for the degree of Doctor of Philosophy of

The Australian National University

David Kernot

October 2018

STATEMENT OF ORIGINALITY

The work presented in this thesis is, to the best of my knowledge, my own original work, except where acknowledged in the text.

David Kernot

Acknowledgements

For someone who left school at 16 to become a boy soldier, the degree of Doctor of Philosophy is the penultimate celebration of what can happen if one applies themselves. This opportunity, to study at The Australian National University's National Security College, would not have been possible without the support of my two academic supervisors, Professor Terry Bossomaier and Professor Roger Bradbury. I am grateful for this opportunity beyond words, and to observe such insightful melding of their collective minds was staggering. It shows in the methods used in the thesis, in their knowledge of complex systems, and their background in Ecology and Biology and the natural sciences. None of this would have been possible without them, and I thank them.

A special mention to my DST Group work colleague, Ph.D. mentor, and supervisor, Dr. David Crone who provided me a solid thesis structure to move on with, along with encouragement and a gently needed shove whenever I began to lag.

ANU's Shakespearean expert, Dr. Kate Flaherty was instrumental at the initial research stages in highlighting that the early findings reflected expert scholars' thoughts on the contested works of Shakespeare.

This research would not be possible without the support of the Defence Science and Technology Group, the Australian Government's lead agency dedicated to providing science and technology support for the country's defence and security needs. I am particularly grateful for the insightful advice provided by Dr David Crone, Dr Coen van Antwerpen, Dr Richard Taylor, Dr Tim Pattison, Dr Aaron Ceglar, Dr Dmitri Kamenetsky, Dr Richard Davis, Dr Yi Yue, Dr Paul Gaertner, and the numerous unknown reviewers, all who contributed to the quality of my eight research papers. This research is also supported by an Australian Government Research Training Program (RTP) Scholarship.

To mum and dad, thank you. And finally, to my wife Olivia, who has supported me throughout every step of this journey, you are amazing.

Abstract

This research project's end goal is on the Lone Wolf Terrorist. The project uses an exploratory approach to the self-radicalisation problem by creating a stylistic fingerprint of a person's personality, or self, from subtle characteristics hidden in a person's writing style. It separates the identity of one person from another based on their writing style. It also separates the writings of suicide attackers from 'normal' bloggers by critical slowing down; a dynamical property used to develop early warning signs of tipping points. It identifies changes in a person's moods, or shifts from one state to another, that might indicate a tipping point for self-radicalisation.

Research into authorship identity using personality is a relatively new area in the field of neurolinguistics. There are very few methods that model how an individual's cognitive functions present themselves in writing. Here, we develop a novel algorithm, RPAS, which draws on cognitive functions such as aging, sensory processing, abstract or concrete thinking through referential activity emotional experiences, and a person's internal gender for identity. We use well-known techniques such as Principal Component Analysis, Linear Discriminant Analysis, and the Vector Space Method to cluster multiple anonymous-authored works. Here we use a new approach, using seriation with noise to separate subtle features in individuals. We conduct time series analysis using modified variants of 1-lag autocorrelation and the coefficient of skewness, two statistical metrics that change near a tipping point, to track serious life events in an individual through cognitive linguistic markers.

In our journey of discovery, we uncover secrets about the Elizabethan playwrights hidden for over 400 years. We uncover markers for depression and anxiety in modern-day writers and identify linguistic cues for Alzheimer's disease much earlier than other studies using sensory processing. In using these techniques on the Lone Wolf, we can separate their writing style used before their attacks that differs from other writing.

Contents

Acknowledgements ..... **iii**
Abstract ..... **iv**
Contents ..... **v**
List of Figures ..... **viii**
List of Tables ..... **x**
List of Equations ..... **xii**
Abbreviations ..... **xiii**
Introduction **1**
1.1 Thesis Statement 1
1.2 Introduction 2
1.3 Research Question 7
1.4 Research Hypothesis 8
1.5 Aim and Scope 9
1.6 Thesis Outline 12
1.7 Publications 14
Background and Related Work **16**
2.1 Background 16
2.2 New Gender Components 20
2.3 Referential Activity 22
2.4 New Sensory Components 28
2.5 Identification of Self 31
2.6 Extremist's Disorders 33
2.7 Classification Techniques 34
2.8 Philosophical Theory 37
2.9 Summary 38
Methodology **39**
3.1 The Approach 39
3.2 Research Hypothesis Experiments 41
3.3 Reference Lists 43
3.4 Equations 44
3.5 Seriation 53
3.6 LIWC 54
3.7 Mann-Whitney U Testing 55
3.8 Step-wise Multiple Regression Analysis 55
3.9 k-fold Cross-validation 56
3.10 Vector Space Method 56
3.11 Imposters Method 57
3.12 Word Accumulation Curves 57
3.13 Hierarchical Cluster Analysis 59
3.14 Principal Component Analysis 59
3.15 Linear Discriminant Analysis 59
3.16 Receiver Operating Characterisation (ROC) Curves 59

3.17 Summary 60
Part One: Elizabethan Authorship Studies 61
Shakespeare, Marlowe, and Cary 62
4.1 Introduction 62
4.2 Material and methods 67
4.3 Results 71
4.4 Discussion 80
4.5 Testing the PCA and LDA Concepts on Contemporary Data 82
4.6 Conclusions 84
4.7 Summary 84
Edward III 85
5.1 Introduction 85
5.2 Methodology 86
5.3 Results 88
5.4 Discussion 95
5.5. Conclusion 98
5.6 Summary 98
The Passionate Pilgrim 99
6.1 Introduction 99
6.2 Methodology 101
6.3 Analysis 102
6.4 Discussion 107
6.5 Conclusion 108
6.6 Summary 109
Part Two: Longitudinal Studies 110
The Dark Lady 112
7.1 Introduction 112
7.2 Methodology 114
7.3 Analysis 114
7.4 Discussion 121
7.5 Conclusion 121
7.6 Summary 124
Language Markers Using POS Analysis 125
8.1 Introduction 125
8.2 Methodology 127
8.3 Analysis 129
8.4 Discussion 140
8.5 Conclusion 143
8.6 Summary 143
RPAS Over Time and CSD Indicators 145
9.1 Introduction 146
9.2 Methodology 148
9.3 Analysis 149
9.4 Discussion 156
9.5 Conclusion 158
9.6 Summary 158
Part Three: Terrorist Characterisation 160
Suicide Attackers 161
10.1 Introduction 161

10.2 Methodology 163
10.3 Analysis 165
10.4 Conclusion 176
10.5 Summary 177
Discussion and Conclusions 178
11.1 Introduction 178
11.2 Significant and Original Outcomes 179
11.3 Answering the Research Questions 183
11.4 Impact 185
11.5 Limitations 187
11.6 Conclusions 188
11.7 Future Research 189
References 192
Appendix A 223
External Data 260

List of Figures

Figure 1: Islamic Caliphate in 750AD 4
Figure 2: Motor and Sensory Regions of the Cerebral Cortex 18
Figure 3: Methodology 40
Figure 4: Multi path problem 54
Figure 5: Word Accumulation Curves. 73
Figure 6: Results of Principal Component Analysis. 77
Figure 7: Results of the Linear Discriminant Analysis. 79
Figure 8: The Venus and Adonis play 79
Figure 9: Iris Murdoch and P.D. James Principal Component Analysis. 83
Figure 10: Linear Discriminant Analysis of Iris Murdoch and P.D. James 83
Figure 11: Edward III gendered Personal pronouns (P) versus Richness (R) 89
Figure 12: Edward III Sensory Adjectives (S) versus Richness (R). 89
Figure 13: Spanish Tragedy to Edward III cosine similarity. 90
Figure 14: Venus and Adonis to Edward III cosine similarity. 91
Figure 15: Imposter Method on Hero and Leander to Edward III. 92
Figure 16: Spanish Tragedy scene comparisons. 93
Figure 17: The Passionate Pilgrim LDA, VSM, and PCA overlays. 103
Figure 18: Richness by Age at Publication 130
Figure 19: Iris Murdoch Content to Function Word Ratio Comparisons. 135
Figure 20: P.D. James Content to Function Word Ratio Comparisons. 136
Figure 21: Iris Murdoch and P.D. James' Sensory Means. 138
Figure 22: Richness (R) by Age for Iris Murdoch and P.D. James 151
Figure 23: Personal Pronouns (P) by Age for Murdoch and James. 152
Figure 24: Referential Activity Power (A) by Age for Murdoch and James. 153
Figure 25: Sensory Adjectives (S) by Age for Iris Murdoch and P.D. James. 154
Figure 26: 1-lag autocorrelation (AR1) for Iris Murdoch and P.D. James. 155
Figure 27: Skewness (G1) for Iris Murdoch and P.D. James 155
Figure 28: LIWC negative emotion and anger categories. 167
Figure 29: ROC curves for Anger and Negative Emotion 168
Figure 30: Stepwise Multiple Regression using RPAV on Suicide Attackers. 170
Figure 31: Classification accuracy for regression scores 170
Figure 32: ROC curves for Anger and RPAV. 171
Figure 33: RPAV stepwise regression results 173
Figure 34: Depression comparison between Murdoch and Suicide Attackers. 175
Figure 35: A map of the cerebral cortex. 182
Figure 36: PCA Scree Plot for Shakespeare, Marlowe and Cary 242
Figure 37: Canonical discriminant analysis of Elizabethan playwrights 244
Figure 38: PtoR plots of Edward III. 245
Figure 39: Iris Murdoch Content Words POS Mean. 249
Figure 40: Iris Murdoch Function Words POS Mean. 249
Figure 41: P.D. James Content Words POS Mean 250
Figure 42: P.D. James Function Words POS Mean 250

Figure 43: Iris Murdoch Visual Sensory Mean....251
Figure 44: Iris Murdoch Auditory Sensory Mean....251
Figure 45: Iris Murdoch Haptic Sensory Mean....252
Figure 46: Iris Murdoch Olfactory Sensory Mean....252
Figure 47: Iris Murdoch Gustatory Sensory Mean....253
Figure 48: P.D. James Visual Sensory Mean....253
Figure 49: P.D. James Auditory Sensory Mean....254
Figure 50: P.D. James Haptic Sensory Mean....254
Figure 51: P.D. James Olfactory Sensory Mean....255
Figure 52: P.D. James Gustatory Sensory Mean....255

List of Tables

Table 1: Outline of different studies 11
Table 2: Chapter Experiments, Aims, and hypotheses 41
Table 3: A summary of all the techniques employed 42
Table 4: Hierarchical Cluster Analysis Membership for 3 clusters 74
Table 5: The Seriation results of Edward III scenes 94
Table 6: OLO seriation results with noise applied. 95
Table 7: Edward III play with the Shakespearian scenes. 97
Table 8: Hamiltonian path distances of The Passionate Pilgrim poems. 105
Table 9: OLO seriation results of The Passionate Pilgrim. 106
Table 10: Dark Lady clustering results 115
Table 11: Rival Poet and Persuasion sonnet clustering results 116
Table 12: TSP seriation of Shakespeare's Dark Lady Sonnets. 116
Table 13: Hamiltonian path distances between the Dark Lady sonnets. 120
Table 14: Iterative Seriation of the Dark Lady sonnets 122
Table 15: Iris Murdoch Mann-Whitney U-test 12-year ranks 131
Table 16: Iris Murdoch Mann-Whitney U-test 12-year statistics 131
Table 17: P.D. James Mann-Whitney U-test 12-year ranks 132
Table 18: P.D. James Mann-Whitney U-test 12-year statistics 132
Table 19: Iris Murdoch Aggregated Content and Function Word Ratios 133
Table 20: P.D. James Aggregated Content and Function Word Ratios 134
Table 21: Function to Content Word Ratios 134
Table 22: Iris Murdoch Mann-Whitney U-test Ranks 136
Table 23: Iris Murdoch Mann-Whitney U-test Statistics 136
Table 24: P.D. James Mann-Whitney U-test Ranks 137
Table 25: P.D. James Mann-Whitney U-test Statistics 137
Table 26: Iterative PCA of Iris Murdoch and P.D. James 139
Table 27: A comparison of the sensory contribution to PCA variance 140
Table 28: Summary of Sensory Means of Murdoch and James. 141
Table 29: Asymptotic Significance from LIWC2015. 166
Table 30: Area under the curve (AUC) for anger and negative emotion. 168
Table 31: Area under the curve (AUC) for anger and RPAV. 171
Table 32: Mann Whitney U-Test comparisons. 172
Table 33: Referential Activity Power data 223
Table 34: Sensory Adjectives data 225
Table 35: Argamon et al.'s (2003) gender study 234
Table 36: Shakespeare, Marlowe and Carey's chunk samples 234
Table 37: The Passionate Pilgrim Poems by Author. 236
Table 38: Pearson correlation coefficient results of RPAS. 237
Table 39: PCA Descriptive Statistics Shakespeare, Marlowe and Cary 238
Table 40: PCA Correlation Matrix for Shakespeare, Marlowe and Cary 239
Table 41: PCA KMO and Bartlett's Test for Shakespeare, Marlowe and Cary 239
Table 42: PCA Communalities for Shakespeare, Marlowe and Cary 240

Table 43: PCA Total Variance for Shakespeare, Marlowe and Cary 240
Table 44: PCA Component Matrix results 241
Table 45: PCA Rotated Component matrix results 241
Table 46: LDA Eigenvalues results 242
Table 47: LDA Wilks' Lambda results 242
Table 48: LDA Discriminant Function coefficients 243
Table 49: LDA Group Centroids 243
Table 50: Iris Murdoch's novels by year published 246
Table 51: P.D. James' novels by year published 247
Table 52: Iris Murdoch Sensory Word Mann-Whitney U-test ranks 247
Table 53: Iris Murdoch Sensory Word Mann-Whitney U-test statistics 248
Table 54: P.D. James Sensory Word Mann-Whitney U-test ranks 248
Table 55: P.D. James Sensory Word Mann-Whitney U-test statistics 248
Table 56: RPAV values of Iris Murdoch's novels 256
Table 57: RPAV values of P.D. James' novels 257
Table 58: List of Suicide attacker 'modus operandi' 258

List of Equations

Equation 1: Richness 45
Equation 2: Personal Pronouns 49
Equation 3: Referential Activity Power 50
Equation 4: Sensory Adjectives 51
Equation 5: Modified Autocorrelation at lag 1 53
Equation 6: Modified Coefficient of Skewness 53
Equation 7: Mann-Whitney U Testing 55
Equation 8: Cosine Similarity 56
Equation 9: Minmax Similarity 57

Abbreviations

A	Auditory
A	Referential Activity Power
AD	Alzheimer's disease
ADF	Australian Defence Force
ANSF	Afghanistan National Security Force
AR1	1-lag autocorrelation
ASG	Afghan Security Guards
AtoR	Referential Activity Power to Richness
BOW	Bag of Words
CIED	Counter Improvised Explosive Device
CSD	Critical Slowing Down
DOCEX	Document Exploitation
DOMEX	Document and Media Exploitation
DSTO	Defence Science Technology Organisation
G	Gustatory
G1	Fisher-Pearson coefficient of skewness
H	Haptic (tactile)
HUMINT	Human Intelligence
HTT	Human Terrain Team
IED	Improvised Explosive Device
IS	Islamic State
ISAF	International Security Assistance Force
IW	Information Warfare
K	Kinaesthetic
LDA	(Stepwise) Linear Discriminant Analysis
LIWC	Linguistic Inquiry and Word Count
NLP	Neuro Linguistic Programming

O	Olfactory
P	Personal Pronouns (Gender)
PCA	Principal Component Analysis
PRS	Primary or Preferred Representational System
PtoR	Personal Pronouns to Richness
R	Richness
RPAS	R: Richness (R), P: Personal Pronouns, A: Referential Activity
	Power, S: Sensory Adjectives
RPAV	R: Richness, P: Personal Pronouns, A: Referential Activity
	Power, and V: Visual from the Sensory Adjectives (S) element
S	Sensory Adjectives
SNA	Social Network Analysis
SPSS	Statistical Package for the Social Sciences
StoR	Sensory Adjectives to Richness
SVM	Support Vector Machine
TTR	Type Token Ratio
VAKOG	Visual, Auditory, Kinaesthetic, Olfactory, Gustatory
V	Visual
VSM	Vector Space Method
UNAMA	United Nations Assistance Mission in Afghanistan

Introduction

1.1 Thesis Statement

The intent of this thesis is to create a method that extracts key linguistic features, or attributes, from a person's writing style or speech that can characterise self for identification purposes and track changes in people's mindset over time due to life events. By collecting open source data within cyberspace, such as blogs and chat room discussions, this identity signature of a person could be used to predict the likelihood of conflict and provide early warning indicators to aid in the defence and regional security of Australia. It might be possible to use this to indicate a tipping point and predict self-radicalisation in people.

To support an automated early warning system that can be deployed in the field, an algorithm must be constructed that can analyse an unknown author's signature from anonymous, unstructured open source texts and through analysis of the content, key stylistic features that describe self must be extracted. By uncovering self with cognitive functions such as aging through word richness, internal gender identification from personal pronouns (Kernot, 2016), abstract or concrete thinking through referential activity emotional experiences (Murphy, Maskit, & Bucci, 2015), and the mind's sensory processing, a signature of a person can be created from their writing style. Visualisation and trend analysis of those key characteristics, messages tagged by richness, gender, referential activity and sensory style can be used to highlight different levels of conflict within that person for early warning security.

This research has its genesis in a Master of Philosophy thesis completed in 2013, where observations about the application of gender and the Representational System sensory-based writing styles highlighted an opportunity to extend its application into the conflict and early warning domain and apply it to current, real-world problems within the national security space.

1.2 Introduction

There have been more than 200 wars since the start of the 20th century, leading to about 35 million battle deaths. However, efforts at forecasting conflicts have so far performed poorly for lack of fine-grained and comprehensive measures of geopolitical tensions (Chadefaux, 2014). After the Vietnam War ended in 1975, the Australian government's military conflicts had been dominated by peacekeeping and low-risk deployments until the ANZUS Treaty was invoked in 2001 after the 9-11 terrorist attacks in the United States (US). Since 2001, Australian military force members and personnel from the International Security Assistance Force (ISAF) and the United Nations Assistance Mission in Afghanistan (UNAMA) suffered casualties in the war in Afghanistan. A significant threat that grew in Afghanistan during this period was the risk of the insider threat (Bordin, 2011; Smith, 2012), the trusted insider, known in military circles as a 'green-on-blue' incident when the individual turns on his friends and colleagues. Insurgents dressed as Afghanistan National Security Force (ANSF) members and Afghan Security Guards (ASG) perpetrated these "green-on-blue" incidents and accounted for a growing trend in that theatre of war. Added to this is the threat from Improvised Explosive Devices (IEDs) placed by unknown insurgents, which are a major cause of death and battle casualty. Such casualties are occurring, partly, because analysts do not have the right tools to identify the perpetrators pre-emptively. This need for tools to help identify perpetrators by scanning textual data and matching identifiable characteristics grew.

Existing tools might help to identify perpetrators by scanning textual data and matching identifiable characteristics. However, these tools only work when characteristics of possible perpetrators are *known*. Exploitation of documents and text (DOCEX) and media (DOMEX) found near insurgents and in hidden caches only help when the perpetrators can be identified. Therefore, tools are needed where the characteristics of the perpetrators are *not* known and can identify them through their linguistic style. A 2010 report by the US highlights that DOMEX is the new intelligence discipline and it crosses all aspects of intelligence (Cox, 2010).

Identifying perpetrators is difficult in unconventional, or asymmetric warfare because the perpetrators hide among the general population. Osama bin Laden, leader of the Islamic terrorist group Al-Qaeda, was reported to have been killed in 2001 after a battle with the US in the Afghan mountain stronghold of Tora Bora, and there were doubts about the authenticity of the 30-35 audio, video, and electronic texts that circulated

(Bloxham, 2011; Cooper, 2011; Griffin, 2009, 2013). Nordin Mohammad Top, a key member of the Al-Qaeda-linked militant group, Jemaah Islamiah rose to prominence after the Bali bombings of 2002, and he was Indonesia's most wanted Islamist militant. In a similar style to bin Laden, he is thought to have escaped a safe-house raid, and doubt was raised about his identity in video footage because bin Laden hid behind a mask. He too was reluctant to use mobile phones, relied on a close network of sympathisers, and used couriers to send messages (Reuters, 2009; Shears, 2009). Many perpetrators operate inside of chat rooms, hidden in cyberspace where they can be anonymous and incite or plan violent acts. They can generate *social influence,* media "buzz," and cause violent responses from their posts (Colbaugh & Glass, 2012). This may have a lot to do with the power of cyberspace because of the anonymity it provides. A recent study has shown that people are more likely to open up to a stranger they meet on the Internet and divulge information about themselves they wouldn't admit to their closest friends and relatives. This confiding of our true-self presents opportunities for people to make friends with someone they have never met (Bargh, McKenna & Fitzsimons, 2002).

The role in Afghanistan has now reduced in response to the diminished threat from the Taliban and Al-Qaeda. Many of Australia's military have returned home from Afghanistan since the major withdrawal in December 2013, when the Afghan people were deemed to be able to manage the threat of the Taliban and Al-Qaeda in the Uruzgan province. However, a new threat has emerged from Al-Qaeda in Iraq[1] in the form of the Islamic State (IS), otherwise known as DAISH (Arab acronym for *Al-Dawlah Al-Islamiyah fe Al-Iraq wa Al-Sham* meaning the Islamic State of Iraq and Syria, or Sham*)* (Saikal, 2015). This new threat is one where the people of Iraq and Syria now face shocking casualties as DAISH fights to create an Islamic caliphate spanning Iraq, Syria[2] and other parts of the Levant to regain the territory they were promised before the Sykes-Picard Agreement of 1916, at the end of the First World War.

The map, (Figure 1), shows the extent of the Islamic Caliphate in 750AD in comparison to the East Roman Byzantine Empire. The three shaded areas highlight the conquests of the Arabs, or Saracens, up to the death of Mohammed in 632AD, then under the first three Khalifs in 632-656AD, and finally the Ommiad Khalifs in 661-750AD (Shepherd, 1926). The Sykes-Picard Agreement was a deliberate attempt by England, France, and

[1] Listed as a Terrorist Organisation in Australia on 2 March 2005 as Tanzim Qa'idat al-Jihad fi Bilad al-Rafidayn (http://www.nationalsecurity.gov.au/Listedterroristorganisations/Pages/default.aspx)
[2] As at November 2017, the ISIS hold on their de facto Syrian capital, Raqqa has collapsed

Russia to reduce the control held by the defeated Ottoman Empire that grew from the 750AD caliphate.[3]

The threat from Islamic State is different from Al-Qaeda because of their approach. Through the power of social media, the group has been able to reach a vast population of disenfranchised youth and recruit them in their fight within Syria and northern Iraq. They have been able to encourage attacks across the globe. Australian military support has again travelled to Iraq, and now face a more complex IED threat from roadside bombs. There are more considerable differences from the fight in Afghanistan because of the inclusive actions of DAISH. Social media and the Internet provided a readily available opportunity to convince young Westerners and Muslims alike to join in their crusade, and it gave them a powerful tool. Now, many of these young people are returning to their homes, some to Australia. DAISH's online magazine, *Inspire,* recently published an article called "*Make a Bomb in Your Mom's Kitchen*," which was translated into Bahasa by Indonesian jihadists (Ramakrishna, 2014).

Figure 1: Islamic Caliphate in 750AD

To draw a parallel on the implications of people open to violent acts of terrorism, on the 7th of July 2005, in London, the first successful home-grown extremist group carried

[3] "The Califate in 750." From The Historical Atlas by William R. Shepherd, 1926. https://commons.wikimedia.org/wiki/File:Califate_750.jpg

out attacks. Akbarzadeh (2013: 453) highlights that these were "the first successful suicide attacks on Western European soil carried out by four British-born men. The attacks catapulted into the public spotlight questions about the capacity of Muslims to integrate into secular, liberal democracies." Samantha Lewthwaite converted to Islam and married one of the London bombers, Germaine Lindsay, after meeting him in an internet chat room, and, after his death has become one of the world's most wanted terrorist suspects (Kirby, 2007; The Sydney Morning Herald, 2013).

The insider threat, or the threat of the lone wolf or lone actor, has been highlighted through the tragic death of passengers when Malaysian Aircraft flight MH370 vanished without a trace in 2014 with some suggestions the cause had something to do with pilot involvement (Holdaway, 2014). This was further reinforced when the co-pilot of Lufthansa subsidiary Germanwings Flight 4U 9525 crashed the plane into the French Alps killing 150 people in a suicide act (Flottau, 2015). Because of technological trends, easy Internet access can expedite direct action, and some military strategists warn of Fifth Generation Warfare in which lone wolves mount crippling cyber-attacks on national infrastructure or deploy small radiological devices, dirty bombs, against cities (Ramakrishna, 2014).

Australia's military involvement has now grown in response to the global threat against terrorism and is involved in over seven military operations throughout the Middle East and North Africa. However, there is a growing threat within Australia and from its near region.

There have been a number of self-radicalised people, lone wolves, conducting terror attacks throughout the world. The Westminster attacker, Khalid Masood, born Adrian Russell Ajao in Dartford, Kent, UK, had converted to Islam. With a violent criminal record, he had been investigated for violent extremism, but he was a 'peripheral figure' and not part of the current intelligence picture. There was no prior intelligence of his intent, or of the plot when he killed five people and injured more than 50 (Bourke & Miller, 2017). Omar Mateen killed 49 people and injured 53 in a lone wolf attack on the Pulse nightclub in Orlando in 2016 and is the deadliest terrorist attack in the United States since 9/11[4], and while officials had no warning. Mateen was repeatedly investigated by the FBI (Byman, 2016). A 15-year-old high school student Farhad Khalil Mohammad Jabar who shot dead a NSW police employee outside the force's

[4] Until 1st of October, 2017 when 64-year-old Stephen Paddock killed 58 people and injured 851 (452 from gunshot wounds) on the Las Vegas Strip in Nevada (https://en.wikipedia.org/wiki/2017_Las_Vegas_shooting)

headquarters in 2015 had not come to the attention of counter-terrorism police before he carried out the "politically motivated" attack, and it is believed the teenager was acting alone (Ralston, Benny-Morrison& Olding, 2015). In 2014, Lindt café gunman, Man Haron Monis, born Mohammed Hassan Manteghi Borujerdi, held 18 people hostage and as a result, two people were killed, but he was not believed to be a threat to national security. There were no indications he intended to engage in terrorism, although he was known to Australian and overseas security agencies for his ongoing offensive and nuisance behaviour with the potential to incite others to violence (Tucker & Gelineau, 2016). Melbourne teenager Numan Haider stabbed two police officers during an arranged meeting outside the Endeavour Hills police station in Victoria in 2014, just days after having undergone a degree of radicalisation (Cooper, 2016). Reinforcing the earlier conclusions of Sarma (2017), here too, many of these individuals were known to authorities and not considered a current threat, but some were not known at all.

There is no universally agreed definition of radicalisation. It is mostly described as the process/es whereby individuals or groups come to approve and participate in violence for political ends, and the UK government's counter-terrorism strategy suggests it occurs when people turn to violence to resolve perceived grievances caused by experiences and events in their life (Stevens, 2009). While examples of self or auto-radicalisation through the internet are rare, the functioning of Web 2.0 facilitates the radicalisation of youth with and without prior inclination toward jihadist activity (Conway & McInerney, 2008). While a lone-actor undergoes their own ideological radicalisation process where personality traits are relevant (Bakker & Roy, 2015). In their exploratory study of YouTube posts, Conway and McInerney (2008), highlight online supporters come from a younger male demographic (25 years and less) outside of the Middle East and North Africa in countries with half originating from the USA (35%) and UK (17%) and 15% from Canada (8%), Australia (4%), and Germany (3%).

There is a consensus that there should be a way to monitor the environment for early warning indications of conflict as highlighted, for example, in the Defence White Paper (Commonwealth of Australia, 2013). This was reinforced in the Defence Intelligence Organisation's web page on intelligence (Commonwealth of Australia, 2002-14). It was also mentioned in the Defence Science and Technology Organisation's 2014 vision of cyber (Commonwealth of Australia, 2014: 17) and with the Australian government's

recent measures on countering violent extremism (CVE) (Commonwealth of Australia, 2015).

The program of research reported in this thesis focuses on the lack of authorship analysis tools and the identification of anonymous authors to identify people that might be insurgents, 'insiders', or a lone wolf. It draws on existing anonymous authorship identification research and techniques (Kernot, 2013) and extends that research by addressing the gaps that were identified.

This research extends these concepts by describing self (for a more detailed determination of self and how it relates to psychology and identity refer to Chapter 2, Section 2.5), and the anonymous identification of authors in cyberspace, by drawing on neuropsychology and neuroscience markers within the brain that appears in writing and discourse analysis. The premise of the research is that by characterising an author's identity using multivariate features, or markers within writing, broken up into four categories, **RPAS**: Richness (R), Personal Pronouns that indicate Gender (P), Referential Activity Power (A), and Sensory-based Adjectives (S), it is possible to identify self from the textual data in cyberspace. Further, we draw on a phenomenon called Critical Slowing Down (CSD), a dynamical property used to develop early warning signs of tipping points (Slater, 2013). By conducting time series analysis using modified variants of 1-lag autocorrelation and the coefficient of skewness, two statistical metrics that increase near a tipping point (Slater, 2013) changes in a person's moods, as they shift from one state to another can be observed. Applying this concept to a lone wolf, before the point when they commit a terrorist act, these hidden characteristics or cues that are inherent in someone's writing style might be able to be used to predict a tipping point and provide early warning signals to prevent a possible crisis.

1.3 Research Question

The objective of this research is to separate the identity of individuals and to highlight changes within them that indicate self-radicalisation. Therefore, the aim of the program of research reported in this thesis is to develop an algorithm that extracts key linguistic features, or attributes, from a person's writing style or speech that can characterise self for identification by data mining open source data within cyberspace. This stylometric identity signature can then be used to predict the likelihood of conflict, in this case,

self-radicalisation within an individual and provide early warning indicators to aid in the defence and regional security of Australia.

Several terms (richness, referential activity, personal pronouns or gender, and sensory adjectives) have been discussed in this chapter. They have been highlighted as potential new stylometric characteristics of language that could identify self and be used in tools to identify people in cyberspace.

Given the need to identify linguistic characteristics of people, several questions can be asked. Can the content analysis of written text extract stylistic features that identify a person? Can a person be described from gender pronouns? Can richness be determined from a person's writing style? Can a referential activity score be determined from linguistic particles? Can a sensory-based representational system be determined from adjectives? Can changes in these stylistic features identify conflict in an individual prior to it occurring? Can the stylistic features be placed into a framework to predict an event for early warning purposes? In considering these questions, it must be understood that they feed the broader and overriding research objective that drives this thesis. Given the need for tools to identify people in cyberspace who wish to harm Australia and its national interests, the research question is:

> *Can the automated extraction of key linguistic attributes from text-based data identify an author's personality, or self, and be used to predict self-radicalisation*?

1.4 Research Hypothesis

There are four research questions addressed:

1. Can a stylistic fingerprint of a person's personality – their personal signature – reveal their 'identity' from their writing style?

2. Does a person's 'identity' change over time because of life events, such as trauma, depression, and disease, or is it stable?

3. Can the application of techniques visualise the critical slowing down phenomena and identify changes in a person's moods, or shifts from one state to another, that might indicate a tipping point for self-radicalisation?

4. Can the final writings of suicide attackers be separated from 'normal' bloggers?

These research questions can be expressed as hypotheses tests, as follows:

Null hypothesis – H_0: The stylistic fingerprint of a person's personality – their personal signature – cannot reveal their 'identity' from their writing style.

Alternate hypothesis - H_1: The stylistic fingerprint of a person's personality – their personal signature – can reveal their 'identity' from their writing style.

Alternate hypothesis - H_2:A person's 'identity' changes over time because of life events, such as trauma, depression, and disease.

Alternate hypothesis - H_3:The application of techniques to visualise the critical slowing down phenomena can identify changes in a person's moods, or shifts from one state to another, that might indicate a tipping point for self-radicalisation.

Alternate hypothesis - H_4: The final writings of suicide attackers can be separated from 'normal' bloggers.

If the approach to identity using personality, through RPAS, is successful and it is possible to create a personal signature of individuals (research hypothesis 1), and to separate 'normal' writing from that written before a terrorist attack (research hypothesis 4), when taking into account the 'normal' changes in a person's personal signature over time (research hypothesis 2), it might be possible to use techniques to visualise the critical slowing down phenomena and determine the tipping point where a disenchanted person becomes self-radicalised (research hypothesis 3). If the answer is yes to all four hypotheses, then it might be possible to stop lone wolves before they act.

1.5 Aim and Scope

The aim of the research reported in this thesis is to develop an algorithm (RPAS) that extracts key linguistic features, or attributes, from a person's writing style or speech that can characterise self for identification by data mining open source data within cyberspace. This identity signature can then be used to predict the likelihood of self-radicalisation in an individual, and aid in the defence and regional security of Australia.

To meet the aim of the research reported in this thesis, three research phases that test the development of the algorithm are required. These are described briefly now.

1.5.1 Phase One

Data and Algorithms Development. In this phase, the algorithms are developed to identify self through four feature-sets comprising twelve linguistic features to classify a person through **RPAS**: Richness (R), Personal Pronouns (P), Referential Activity Power (A), and Sensory-based Adjectives (S). Reference data to identify a person is also generated to score the feature-sets.

1.5.2 Phase Two

Experiments. In this phase, seven studies are conducted across three distinct groups or phases to test key aspects of the research question. Table 1 shows the logic and structure of the experiments. Each of the experiments is discussed in the subsequent paragraphs.

Study 1 - Authorship Identification. In this study, the works of William Shakespeare, Christopher Marlowe, and Elizabeth Cary are used to test the premise that RPAS can be used to create signatures of more than one person and identify self.

Study 2 - Authorship Identification. In this study, the works of William Shakespeare, Christopher Marlowe, and Thomas Kyd are used to test the premise that RPAS can be used to identify the authorship of an unknown author's work, through the anonymous play *Edward III*.

Study 3 - Authorship Identification. In this study, the works of William Shakespeare, Christopher Marlowe, Thomas Kyd, Bartholomew Griffin, and Richard Barnfield, are used as a test case to test the premise that RPAS can be used to identify the authorship of many unknown author's works, through the publication, *The Passionate Pilgrim*.

Study 4 - Authorship Changes over Time. In this study, the single work of William Shakespeare, *The Sonnets*, is used to test the premise that RPAS can be used to characterise subtle differences, or changes, in a person's writing style within small texts over time.

Study 5 – Authorship Changes over Time. In this study, the works of Iris Murdoch and P.D. James are used to test if an author's characteristics change over time using RPAS. This time, the study is conducted within larger texts when one author has

depression and Alzheimer's disease that might mimic the proximal events and life stressors faced by a terrorist.

Study 6 – Authorship Changes over Time. In this study, the works of Iris Murdoch and P.D. James are used to test if a tipping point can be discovered using the Critical Slowing Down (CSD) dynamical property prior to when a life-changing event occurs. In this case, the event is Iris Murdoch's Alzheimer's disease progression. However, we believe this mimics the proximal events and life stressors faced by a terrorist as they become self-radicalised or prior to them conducting a terrorist act.

Study 7 – Lone Wolf Study. In this study, the suicide notes and final manifestos of suicide attackers are compared to normal bloggers and a person with depression to see if their writing can be separated using RPAS.

1.5.3 Phase Three

Evaluation and Recommendations. In this phase, the effectiveness of the algorithm to determine self and provide conflict early warning against the research question will be evaluated. Recommendations will be made on further research.

Table 1: Outline of different studies

Study	Chapter	Aim	Data Source
1	4	Identifying features to identify authors (known authorship) from RPAS	Shakespeare, Marlowe, Cary
2	5	Testing RPAS on an unknown author.	Edward III
3	6	Testing RPAS on a set of multiple unknown authors.	Passionate Pilgrim
4	7	Identifying features of a single author to identify characteristics in small texts that change over time from RPAS.	Shakespeare's The Sonnets
5	8	Identifying features of two authors to identify characteristics that change over time from RPAS when one has proximal events and life stressors that mimic a potential terrorist.	Iris Murdoch and P.D. James
6	9	Identifying characteristics that change over time using a tipping point and Critical Slowing Down phenomena from RPAS prior to a life changing event.	Iris Murdoch and P.D. James
7	10	Identifying different RPAS features of lone wolf suicide attackers that are different from normal or depressed writers.	Lone wolf suicide attackers

1.6 Thesis Outline

There are eleven chapters and bibliography in this thesis, and the outline of them is as follows:

- Chapter 1 – Introduction. We highlight the need for research in the field of self-radicalisation and how through author identification using personality or self might be beneficial.

- Chapter 2 – Background and Related Work. A discussion on the changing nature of warfare brought about by the asymmetric tactics employed by insurgents, and summarise key findings of critical earlier research. We show how through several types of analysis of anonymous documents that it may be possible to identify insurgents, terrorists, and the lone wolf. The chapter focuses predominantly on examining the methods to identify a person and uncover self. We describe how word richness, a person's internal gender from personal pronouns, Referential Activity, and the sensory Representational System might overcome some of the current anonymous author identification shortfalls. In this chapter, we also state the research hypotheses that stem from the research objectives and questions.

- Chapter 3 – Methodology. We address phase one, Data and Algorithms Development, and describe the methodology that is best used to gather the data and create the author signatures. We restate the research hypotheses and how they are tested through a series of studies. We describe how the reference data list selection is created and highlight how the works are reduced into a Bag of Words (BOW). We identify the RPAS equations that will be used to generate a signature of an author that describe self.

- Chapter 4 – Authorship Identification. We address one part of Phase Two, Experiments, and describe the first series of experiments (one through three). Drawing on the Elizabethan playwrights, and the works of William Shakespeare, Christopher Marlowe, Elizabeth Cary, Thomas Kyd, Bartholomew Griffin, and Richard Barnfield, the proposed algorithms are tested over three experiments to see if they can identify self and characterise an author's writing style. RPAS is tested on the works of William Shakespeare, Christopher Marlowe, and Elizabeth Cary to see how effective they are at separating the writing of different authors.

- Chapter 5 – Authorship Identification. Again, we address one part of Phase Two, Experiments, and describe the second study in the first series of experiments on authorship identification. Drawing on the works of William Shakespeare, Christopher Marlowe, and Thomas Kyd, the proposed algorithms are tested to see if we can identify known authorship from an author's anonymous writing style through the anonymous play, *Edward III*.

- Chapter 6 – Authorship Identification. Again, we address one part of Phase Two, Experiments, and describe the last of the series of experiments looking at authorship identification. Drawing on the works of William Shakespeare, Christopher Marlowe, Thomas Kyd, Bartholomew Griffin, and Richard Barnfield, the proposed algorithms are tested to see if we can identify known authorship from a more complex dataset when there are multiple unknown authors through the publication, *The Passionate Pilgrim*.

- Chapter 7 – Authorship Changes Over Time Study. We address another part of Phase Two, Experiments, and describe the second series of experiments (four through six). This first study in this next part, we draw on William Shakespeare's work, *The Sonnets.* We examine the changes in writing in a single author over time.

- Chapter 8 – Authorship Changes Over Time Study. Again, we address another part of Phase Two, Experiments, and describe the second experiment looking at authorship changes over time. Drawing on the works of the contemporary authors, Iris Murdoch and P.D. James, we examine the effects of time on an author's signature. We examine larger texts when one author has depression and Alzheimer's disease that might mimic the proximal events and life stressors faced by a terrorist.

- Chapter 9 – Authorship Changes Over Time Study. Again, we address another part of Phase Two, Experiments, and describe the third and final experiment looking at authorship changes over time (study six). Again, the works of the contemporary authors, Iris Murdoch and P.D. James are examined, but this time they are used to test if a tipping point can be discovered using the Critical Slowing Down (CSD) dynamical property prior to a life-changing event occurring that mimics the proximal events and life stressors faced by a terrorist as they become self-radicalised or prior to them conducting a terrorist act.

- Chapter 10 – Lone Wolf Study. We address the final part of Phase Two, Experiments, and describe the third series of experiments (study seven). In this study, we compare the suicide notes and final manifestos of suicide attackers to normal bloggers and a person with depression to see if their writing can be separated.

- Chapter 11 – Discussion and Conclusions. We address Phase Three, Evaluation and Recommendations, and demonstrate that the aims as stated in Chapter One have been achieved by meeting the research objectives. We describe the significance of our findings and its limitations and provide our conclusions and where future research in this domain should be directed.

1.7 Publications

- From Chapter 4: **Kernot, D**., Bossomaier, T., & Bradbury, R. (2018). *Shakespeare's Sotto Voce: Determining True Identity from Text. Frontiers in Psychology* Vol 9. March 2018 Article 289, 1-17.

- From Chapter 5: **Kernot, D**., Bossomaier, T., & Bradbury, R. (2017). *Did William Shakespeare and Thomas Kyd Write Edward III?* International Journal on Natural Language Computing. Vol. 6, No. 6. December 2017.

- From Chapter 6: **Kernot, D**., Bossomaier, T., & Bradbury, R. (2017). *Stylometric Techniques for Multiple Author Clustering: Shakespeare's Authorship in The Passionate Pilgrim.* International Journal of Advanced Computer Science and Applications. Vol. 8 No. 3, 1-8.

- From Chapter 7: **Kernot, D**., Bossomaier, T., & Bradbury, R. *(2017). Novel Text Analysis for Investigating Personality: Identifying the Dark Lady in Shakespeare's Sonnets*. Journal of Quantitative Linguistics. Vol 24 No 4, 255-272.

- From Chapter 8: **Kernot, D**., Bossomaier, T. and Bradbury, R. (2017). *The Impact of Depression and Apathy on Sensory Language.* Open Journal of Modern Linguistics, **7**, 8-32.

- From Chapters 8 and 9: **Kernot, D**., Bossomaier, T., & Bradbury, R. (2017). *The Stylometric Effects of Aging and Life Events on Identity*. Journal of Quantitative Linguistics. Published online 6 Dec 2017, 1-21.

- From Chapter 10: **Kernot, D**., Bossomaier, T., & Bradbury, R. (2017). *Identifying Suicide Attackers in Cyberspace* (under review).

- From Chapters 4-10: Bradbury, R., Bossomaier, T., **Kernot, D**. (2017). *Predicting the emergence of self-radicalisation through social media: A complex systems approach.* In Conway, M., Jarvis, L., Lehane, O., Mcdonald, S., Nouri, L. (eds) Terrorists' Use of the Internet: Assessment and Response. IOS Press. Vol 136, 379-389.

Background and Related Work

In this chapter, we describe the context of the work in which this program of research reported in this thesis is situated and provide some background for the work this research draws from. Drawing on the research questions and hypotheses, we highlight the problems faced by Australians in this context and summarise the related work that extends the earlier research.

2.1 Background

The program of research reported in this thesis draws on a Master of Philosophy thesis (Kernot, 2013). The thesis proposed that by being able to create a signature that described an anonymous person from their texts in cyberspace, that insurgents and bomb-makers of Improvised Explosive Devices (IEDs) could be identified and that this, in turn, could be used to reduce the threat to deployed Australian troops stationed within Afghanistan.

2.1.1 Gender Component

In this section, the earlier work in the area of gender is described (Kernot, 2013), which was used to identify an anonymous author from their writing style. The concept drew on Kernot (2013) from an 'author's invariant' (Stanczyk & Cyran, 2007), as a method to detect subtle, hidden characteristics in written text that could differentiate one author from another. Anonymous authorship identification techniques were tested through two novel approaches. The first approach was based on the assumption that gender could be defined by a person's use of personal pronouns. The study drew on the use of pronouns in gender identification (Argamon *et al.,* 2003; Argamon *et al.,* 2007; Chung & Pennebaker, 2007; Harré, 1999a, 1999b; Hota *et al.,* 2006; Koppel *et al.,* 2002; McGrath, 2003; Newman *et al.,* 2003; de Vel *et al.,* 2001), and in particular the use of gender-based pronouns to determine authorship, by Argamon, Koppel, Fine, and Shimoni (2003). More recent work had been done by others (Herring & Paolillo, 2006; Kagstrom *et al.,* 2009; Lai, 2009) but did not achieve the gender classification accuracies greater than Argamon *et al.*'*s* (2003) level of 80%.

By extending Argamon *et al.'s* (2003) study and using their initial list of gender-based pronouns, Kernot (2013) assigned a gender to each pronoun, labelled M (more likely used by a male gender) or F (more likely used by a female gender) based on Argamon *et al.'s* (2003) median results, T-test significance levels, male and female means, and standard errors. This new reference list was compared against a Bag of Words that comprised 30 samples taken from the Internet, with an average size of 1,000 words each (6 articles from 5 authors with a contribution of 6,000 words per author) resulting in a total Bag of Words size of 30,000 words. This list was reduced to 1,981 instances of the gender-based pronouns. Using logistic regression analysis on the data, 27 of the 30 samples (90%) were correctly identified by gender using the three most statistically significantly words (*my*, *her*, and *its*), and a gender formula drew on the logistic regression predictor variables and coefficients.

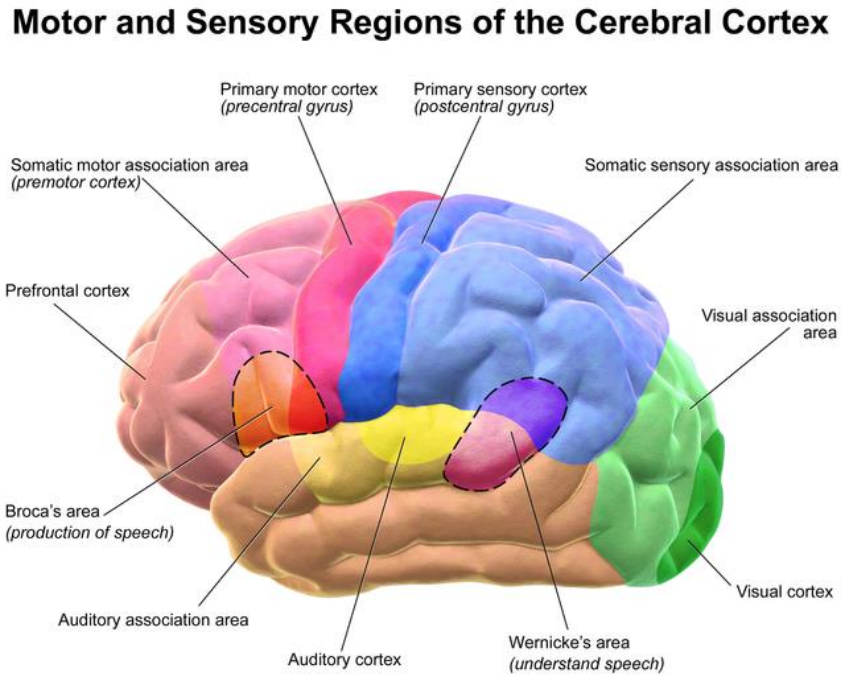
A recommendation was made that the algorithm is further tested against wider and larger datasets. There was a belief that, outside of defining gender as male or female, the gender scores could be used to provide more fidelity to an author's anonymous signature.

2.1.2 Sensory Component

In this section, we describe the earlier work in the area of Neuro-Linguistic Programming (NLP in this case does not refer to the commonly referred term Natural language Programming) sensory predicates (Kernot, 2013), which was used to identify an anonymous author from their writing style. The concept was based on the assumption that through sensory predicates (nouns, verbs, adverbs, and adjectives), it was possible to identify a person from the way they used their five senses and the different part of their sensory cortices. The term VAKOG (V – visual, A – auditory, K – kinaesthetic, O – olfactory, G – gustatory), described the Representation System (RS), which is the part of the brain that processes our five senses of sight, hearing, touch and body sensations, smell, and taste (see While cognitive psychology is more focused on the mind, cognitive neuroscience focussed on the brain (Bechtel, 2002). Here, we use the sensory processing of the brain and how that impacts on the psychological aspects of the mind through linguistics. Our interests lie in how basic sensory information is represented, and how internal and external sensations of prior experiences are stored (see Oosterwijk et *al*., 2015) and integrated into language and cognition through a network of overlaid and combined neural functions, such as auditory, visual, tactile, and sensory-motor sensations (St Clair, 2017).

Figure 2)[5]. While cognitive psychology is more focused on the mind, cognitive neuroscience focussed on the brain (Bechtel, 2002). Here, we use the sensory processing of the brain and how that impacts on the psychological aspects of the mind through linguistics. Our interests lie in how basic sensory information is represented, and how internal and external sensations of prior experiences are stored (see Oosterwijk et *al*., 2015) and integrated into language and cognition through a network of overlaid and combined neural functions, such as auditory, visual, tactile, and sensory-motor sensations (St Clair, 2017).

Figure 2: Motor and Sensory Regions of the Cerebral Cortex



The sensory modalities of the Representation System was grounded in Neuro-Linguistic Programming (NLP) techniques, the study of the structure of human subjectivity, and a model to demonstrate the basic process used to encode, transfer, guide, and modify our behaviour (Bandler & Grinder, 1979; Grinder & Bandler, 1976; Dilts *et al*., 1980; Grosu *et al*., 2017; Marashi & Abedi, 2017). It grew from an idea that individuals have a "private language," a world of sensations, that is unique and individual to that person, and it defined how people interpreted their environment, and how they attached a 'feeling' to each word expressed through their senses

[5] From Blausen.com staff (2014). "Medical gallery of Blausen Medical 2014". *WikiJournal of Medicine* **1** (2). DOI:10.15347/wjm/2014.010. ISSN 2002-4436.

(Wittgenstien, 1922). The Primary Representational System (PRS), a preferred or primary sensory cortex that people draw on to speak and write (Grinder & Bandler, 1976) unconsciously determines a person's use of noun/verb/adverb predicates in their communication, and this is reflected in both their oral and written communication style.

There was an assumption that NLP could provide techniques by which these hidden cues could be unmasked as a person speaks and/or writes because each person has a PRS (Grinder & Bandler, 1976; Heap 2008; Mercia & Johnson, 1984; Skinner & Stephens, 2003).

After extensive debate on the scientific validity of NLP (Heap, 1988a; Heap, 1988b; Heap, 2008; Sharpley, 1984; Sharpley, 1987), it had mostly been abandoned by academia and psychology (Witkowski, 2010). However, Colaco *et al.* (2010) developed a psychometrically-based Neuro-Linguistic analysis tool called NEUROMINER to classify one type of computer programmer from an OSS mailing list content and they used it to identify developer's personalities and general emotional content (Rigby & Hassan, 2007) from sensory-based Visual, Auditory and Kinaesthetic words or phrases.

The Kernot (2013) study drew on the success of the Colaco *et al.* (2010) work and again tested the thirty articles obtained from the Internet by five authors across a range of sources. A sensory-predicate algorithm was developed from discriminate analysis testing, and while no PRS could be identified, it was found that a signature of anonymous individuals could be created from a combination of all of the Representational System (RS) scores to identify people.

The program of research reported in this thesis extends these concepts above by describing self. The focus is on the anonymous identification of authors in cyberspace and characterising them by the way people think. It draws on neuropsychology and neuroscience markers within the brain that appear in writing and discourse analysis. We draw on a Critical Slowing Down (CSD), a dynamical property used to develop early warning signs of tipping points (Slater, 2013), and the use in a person is a relatively new concept (Trefois *et al.*, 2015; Meisel *et al.*, 2015; Scheffer, 2010; Slater, 2013; van de Leemput *et al.*, 2014). It is discussed in more detail below (Section 2.3.3 Critical Slowing Down), but it has the potential to be used to predict the likelihood of events and identify the author of those texts in cyberspace.

2.2 New Gender Components

"There is no general agreement as to what extent – if at all – the psychological make-ups of the two sexes are different by nature, but there is no doubt that gender is represented through a person's choices of lexical and functional items" (Dam, 2014: 87-88). In determining the role of personal pronouns in the construction of female identity, Dam (2014) highlights the use of pronouns 'we', a deictic or dependent upon context pronoun, 'you,' and 'your,' as contributing to identity.

Hancock *et al.* (2014) suggest that recent research on gender differences in language present a diminishing picture. In their study (Hancock *et al.,* 2014) of whether language use can predict perceptions of gender and femininity, 10 males, 12 females, and 13 transgendered women (male-to-female) speakers were rated against femininity. What the study highlighted was that 4 of the 14 variables differentiated males from females using *T*-units (known as the minimal terminable unit, where each T-Unit "is the shortest units into which a piece of discourse can be cut without leaving any sentence fragments and would contain one independent clause and its dependent clauses" (Hunt, 1965: p.189) and in Hancock *et al.* (2014) it focussed on dependent clauses and personal pronouns). Hancock *et al.* (2014) found that transgendered females tended to be more distinct from the females than the men were. In the second part of the study where they compared the use of language, the transgendered women's scores were not significantly different from the men, but none of the 14 individual variables alone were robust predictors of perceived gender or femininity.

In a gender study, Lenard (2014), analysed 204 random speeches from the 113th United States Congress, split equally by gender with the LIWC (Linguistic Inquiry and Word Count) tool from Pennebaker *et al.* (2001). Lenard (2014) used US Congress speeches and analysed 70 language categories and highlighted that women's personal pronoun use was larger than men and they used *the word* '*we*' more. Men used nouns, articles, and numbers more and tended to use the pronoun '*I*' more.

What is clear is from a literature search for further authorship identification techniques in the area of gender, that pronouns are a crucial part of speech in determining this.

2.2.1 Key Pronoun Studies

Flekova and Gurevych (2013), identified people by age and gender on social media websites using a combination of features, which included pronoun ratio for age, but

ignored pronouns to determine gender, and achieved a gender match of 0.58. Rangel and Rosso (2013) drew from the work of Pennebaker (2011) and used a Support Vector Machine method to categorize Parts Of Speech (POS), but this performed better at identifying age than gender. They achieved a gender match of 0.57 when including singular and plural pronouns into their POS list. Argamon *et al.* (2009) used a combination of pronouns (I, me, him, my) to identify females, but achieved 5-10% better success rates when they also considered content words about technology (males) and personal life and experiences (female). They achieved a gender match of 0.76 when looking at style and content feature sets. In Argamon *et al.* (2007), a study of how writing topic and style vary with age and gender of the blogger highlighted similar gender results to the Argamon *et al.* (2003) study. That Articles and Prepositions are used significantly more by male bloggers, while personal pronouns, conjunctions, and auxiliary verbs are used significantly more by female bloggers.

From the literature, much of the work that includes gender has focussed on determining age and looking at author emotion. In all cases, the focus of determining gender was to categorise people as male or female, and this was done using many different techniques with varying success. The closest aligned work was from Argamon *et al.* (2009), and this was the study with the highest success rate of gender matching at 76% from female use of the pronouns 'I,' 'me,' 'him,' and 'my.' While this is not as high as the 80% reached from Argamon *et al.* (2003), it demonstrates a reduction of the key contributing pronouns for gender identification and the power of personal pronouns. In the Kernot (2013) study, a gender success rate of 90% was reached using three pronouns, 'my,' 'her,' and 'its.' A score of 93.3% was achieved using five pronouns, 'my,' 'her,' 'its,' 'themselves,' and 'them,' but the underlying statistical results were not as significant, or reliable. Cheng *et al.* (2011) make the distinction between biological sex as male and female, and a person's gender-related language as a socially constructed aspect where not all men are masculine and not all women are feminine, but internally they can be other than their biological sex.

2.2.2 Summary

What is important here is, outside of the categorisation of people into purely male and female, the existing work in Kernot (2013) provides an opportunity to describe a person on a continuum between 0 and 1 from their use of personal pronouns for identification. At one end, they are likely to be female, and at the other male, but the scores from the gender equation (see Equation 2) provide an opportunity to identify

the internal, socially constructed gender of a person. Another important point is the power of a pronoun to relate to self closely, and this point is researched further and discussed next in Referential Activity.

2.3 Referential Activity

It became apparent that the pronoun is a key part of speech and it is able to identify a person from gender (refer to the section above). The use of first person pronouns is also associated with the identification of age, sex, depression, illness, and more broadly, self-focus (Pennebaker & Lay, 2002), and there are other ways to use pronouns outside of gender to identify self. Pronouns are relatively easy to identify cross-linguistically (Simon & Wiese, 2002), in other languages other than English which highlights their importance in language construction and their value to self.

In the *Psychological Aspects of Natural Language Use: Our Words, Our Selves*, Pennebaker, Mehl, and Niederhoffer (2003) move us from the single pronoun, through the function word type of particles, and on to the concept of Referential Activity, defined as the function of connecting non-verbal experience with language (Mergenthaler & Bucci, 1999), and underpinned by multiple code theory, a "cognitive model that emphasises the role of emotion in human cognition and the complex issues involved in translating emotional experience to verbal form." (Bucci, 1997: 153). The five key areas relevant to this thesis are listed below:

(a) The words people use in their daily lives can reveal important aspects of their social and psychological worlds.

(b) Pronouns may be an overlooked linguistic dimension and are markers of self versus group.

(c) Pronouns are a group of function words known as particles, the glue that holds nouns and regular verbs together, and they can serve as markers of emotional state, social identity, and cognitive styles. Particles are processed in different regions of the brain and in different ways than content words.

(d) Particles include pronouns, articles, prepositions, conjunctives, and auxiliary verbs. There are fewer than 200 commonly used particles, and they account for over half of the words we use.

(e) Particles are referential words that have tremendous social and psychological meaning. In particular, third person pronouns and prepositions capture the ability to verbalize nonverbal experiences through Referential Activity.

This self-referential process concerns stimuli that are experienced as strongly related to one's own person (Northoff *et al.,* 2006). Another way to describe it would be *self*, and Northoff *et al.* (2006: 441) highlight that "When objects and events are viewed through the eyes of the self, stimuli are no longer simply objective aspects of the world, but they typically become emotionally coloured, and thereby more intimately, related to one's sense of self."

Grounded in Critical Realism, an approach that asserts the independence of an external world whilst accepting that knowledge of that world is socially constructed and transient (Bosward, 2017), the American philosopher, Roy Wood Sellers provided a linguistic framework guided by the brain's sensory referential sensations (Sellars, 1916; 1959; 1961). Sellars highlighted referential activity as a biological mechanism that had 'higher levels.' The concept was picked up for clinical studies into depression (Bucci & Freedman, 1981; Bucci, 1982; Bucci, 1984; Bucci & Millar, 1993).

Wilma Bucci (2002) suggests we see all things through the lens of memory schemas. Grounded in neuropsychology from the levels of awareness and the sense of self by Damasio (2003), Bucci draws on multiple code theory (Bucci, 1997), to highlight that humans represent and process information through symbolic codes, and *sub-symbolic* or *non-symbolic codes*, to link words to non-verbal representations through a system that sits between the words (*symbolic codes)* and the subordinate non-verbal modalities (*sub-symbolic* or *non-symbolic codes*). Within the symbolic codes, there are non-verbal symbolic codes, or modalities (multi-modal images within the brain are fused within the posterior parietal region, and to a lesser extent the anterior parietal regions of the brain (see Kayser & Shams, 2015; Andersen & Gnadt, 1989) and refer to somatic sensory association area in While cognitive psychology is more focused on the mind, cognitive neuroscience focussed on the brain (Bechtel, 2002). Here, we use the sensory processing of the brain and how that impacts on the psychological aspects of the mind through linguistics. Our interests lie in how basic sensory information is represented, and how internal and external sensations of prior experiences are stored (see Oosterwijk et *al*., 2015) and integrated into language and cognition through a network

of overlaid and combined neural functions, such as auditory, visual, tactile, and sensory-motor sensations (St Clair, 2017).

Figure 2), and verbal symbolic codes, or modalities (words or low level symbols within the mind).

A structure exists where discrete word entities refer to other discrete words and images, and similarly, discrete images refer to other entities across all of the sensory modalities (visual, auditory, tactual, kinesthetic, olfactory, and gustatory representing what we see, hear, feel, smell, and taste). With the sub-symbolic system, there is systematic processing that occurs across each sensory modality as sounds, smells, feelings, not as words and images, but through representations at the modality level. The sub-symbolic representations are only expressed indirectly by abstract symbols of the verbal symbolic codes, and these discrete representational elements, or non-verbal symbols, connect to the symbols of the verbal symbolic codes. In Kosslyn (1994) there is an emphasis on the notion that information is represented as images and in Kosslyn (2005) a suggestion that this mental imagery draws on many of the same mechanisms as visual perception.

Referential Activity (RA) is grounded in experimental cognitive psychology (Bucci & Kabasakalian-McKay, 2004: 3), and is defined as: "activity of the system of referential connections between verbal and non-verbal representations, as reflected in language style. Nonverbal representations include imagery in all sense modalities, as well as representations of action, emotion, and somatic experience…"

Clinical psychologists use the psycholinguistic variable, Referential Activity to score a person from their speech across four categories (Bucci, & Kabasakalian-McKay, 2004; Murphy, Maskit & Bucci, 2015): the extent that verbal expressions refer to sensate properties of actual things or events or to anything that is experienced as a sensation or feeling sensory characteristics of language (Concreteness); the vividness and effectiveness with which the speaker's language is reflecting and capturing imagery or emotional experience, in any sense modality (Imagery); the quality of detail, e.g. degrees of articulation (Specificity); and, the organisation and focus (Clarity). The RA measures the degree to which a speaker or writer is able to translate experiences into words in a way that evokes corresponding experiences for the listener or reader.

"The concepts of multiple code theory and the referential process is central to both consciousness and the sense of self." (Bucci, 2002: 766). Over a period of many years,

the human approach to coding RA has been automated and tested to perform as well, if not better than humans. The Computerized Referential Activity (CRA) program was developed to model against human raters of RA (Mergenthaler & Bucci, 1999) and from it grew the Weighted Referential Activity Dictionary (WRAD) dictionary of psycholinguistic variables which was used through the Discourse Attributes Analysis Program (DAAP) to automate the human process of determining a person's RA score (Bucci & Maskit, 2004; Bucci & Maskit, 2006; Bucci, Maskit, & Murphy, 2015). But there is another source of computerised psycholinguistic data. The Medical Research Council (MRC) Psycholinguistic Database is a computerised database of psycholinguistic information with around 98,538 words (Colheart, 1981) and it has a number of key linguistic word properties, including scores for imageability and concreteness (from RA).

2.3.1 Pronouns and Depression

Pronouns are important in creating an identity and a sense of self (Priest, 2013). Mental illnesses, such as depression and post-traumatic stress disorder (PTSD), are known to significantly alter pronoun use (Preotiuc-Pietro *et al.*, 2015). The use of first person pronouns in people with depression is clear (Bucci & Freedman, 1981; Weintraub, 1981; Stirman & Pennebaker, 2001; Rude, Gortner, & Pennebaker, 2004; Ramirez-Esparza *et al.*, 2008). Bucci and Freedman investigated the relationship between Referential Activity and depression and found impairment in referential activity with clinical patients with severe depression (Bucci & Freedman, 1981). This may in part be linked because of the contribution that particles, including pronouns, play in rating RA.

2.3.2 RA and Dementia

While some of the tests used to rate RA include spontaneous discourse, when a person is asked a series of non-scripted questions, this requirement of an immediate structured response puts pressure on the cognitive-linguistic system that through linguistic analysis of word use can highlight neurological diseases such as dementia (Bersha *et al.*, 2015). Drawing on a number of studies (Nicholas *et al.*, 1985; Smith, Chenery, & Murdoch, 1989; Holm, Migne, & Ahlse, 1994; Kemper, Thompson, & Marquis, 2001; Maxim, Bryan, & Thompson, 1994; Bird *et al.*, 2000), Bersha *et al.* (2015) reinforce the value of Referential Activity, when they highlight that dementia patients use high frequency low imageability words. They also report on other key linguistic feature that describes self, such as a reduction in available vocabulary and lexical

repetition. This richness can be mapped through the number of unique words used in the written text (Tweedie & Baayen, 1998), and an alternate way of describing it can be through species diversity (Karydis & Tsirtsis, 1996), or Menhinick's Index (Menhinick, 1964).

2.3.3 Critical Slowing Down

It is an important notion that function words (the higher set of words more than particles) do vary according to a person's psychological state (Chung & Pennebaker, 2007). By combining discrete and continuous dynamics a rigorous, expressive, and computationally-tractable framework for modelling the dynamics of the complex, highly-evolved networks can be achieved (Colbaugh & Glass, 2012). Slater (2013) highlight that complex systems across a variety of disciplines (ecology, climatology, finance, and medicine) demonstrate tipping-points or an abrupt, rapid change of state. Ecosystems and biological systems are known to be inherently complex and to exhibit nonlinear dynamics, and changing system dynamics have been suggested as early warning signals (EWS) for tipping points (Trefois *et al.*, 2015). In his study of social media sentiment measuring happiness, anxiety, and tension in blogs, Slater (2013) highlighted that protests and rebellion manifest as tipping points through a rise in anger, sadness or tension. In a comparative study of four emotions in healthy and depressed people, van de Leemput *et al. (*2014) discovered that individuals go through a major transition in moods that are separated by tipping points.

Early warning signals of such tipping-points can be detected via a concept called critical slowing down (CSD), and when measuring variation, 1-lag autocorrelation, or skewness, an increase can be observed near the tipping point. (Slater, 2013; Scheffer, 2010). Slater (2013) highlights that a composite CSD indicator can be computed based on a method introduced by Drake and Griffin (2010) and by using time series analysis, statistical signatures from moving window calculations, can plot the coefficient of variation, 1-lag autocorrelation, and coefficient of skewness. They highlight that to move to predictive analysis, a better understanding of the robustness of the CSD metrics is needed. Drawing on a broad class of neuroscience modelling Meisel *et al.* (2015) suggest an improved estimation of tipping points will occur by incorporating scaling laws. Scheffer (2010) suggests that work in different scientific fields is now suggesting the existence of generic early-warning signals that may indicate for a broad class of systems if a critical threshold is approaching.

Bos and De Jonge (2014) suggest van de Leemput *et al.*'s (2014) results might be correct in their conceptualisation of depression as a dynamic system, but highlight the empirical evidence in the study is weak, and further studies should disaggregate (inter and intra) individual variability more carefully. Wichers *et al.* (2014) further clarified van de Leemput *et al.* (2014) and agreed in part, highlighting that time series assessments obtained while individuals undergo a transition would be ideal because these would allow for direct intra-individual tests. The logic behind their analysis was that, if individuals display early warning signals when closing in on a transition, then individuals who are closer to a tipping point should show higher levels of autocorrelation and variance. In ecosystems, Dakos *et al.* (2012) state that CSD the coefficient of variance approach can be systematically underestimated if the rates of change are slow relative to the frequency characteristics of the forcing regime, and therefore close to the tipping point it might increase or decrease. However, autocorrelation is more effective because it always increases toward critical transitions.

If social media sentiment measuring happiness, anxiety, and tension in blogs can highlight tipping points, manifested as a rise in anger, sadness or tension (Slater, 2013), then we would argue that as emotions such as anxiety, tension, fear and sadness increase prior to a tipping point, that the same change in a person's mindset can be measured through a change in the use of particles (a subset of function words) used to measure Referential Activity. Function words contain emotion like content words. Hancock *et al.* (2007), in a study of 40 dyadic interactions of happiness versus sadness, suggest that people can express emotion through text without any other context, and they discovered happy people wrote larger texts, and while the pronouns used were not different, articles differentiated people's emotions. In people with neurodegenerative disorders, there can be a reduction in their syntactic complexity, in their proportion of words in sentences, and in the proportion of nouns with determiners (Garrard *et al.*, 2005). Determiners are function words, and Pennebaker (2011) highlights function words, pronouns, and articles can indicate the ways people think, feel and connect.

2.3.4 Summary

Three key points arise from this section on Referential Activity. The first is, that if particles can highlight depression and also describe self, and as we have seen, if concreteness and imagery from RA can highlight dementia and depression and reflect self, then by selecting particles from the MRC Psycholinguistic dataset that are high in

concreteness and imageability, it should be possible to identify self through a person's cognitive mental state.

The second point is that there is a strong link between the gender identification pronouns, word particles used in Referential Activity, and our use of sensory words of self through pronouns and word particles, and to verbalising non-verbal ideas and concepts within the sensory cortex that reinforces the sense of self and identity. The third point is that word richness through a species diversity equation can describe a person's unique word choice that reflects self through their cognitive function and contribute to being able to identify self. The fourth point is that by using the concepts mentioned in the points above, it should be possible to measure a tipping point through Critical Slowing Down.

2.4 New Sensory Components

While Neuro-Linguistic Programming has been condemned (Witkowski, 2010), it has been used to enhance transcript analysis (Tosey & Mathison, 2010) in the area of psychophenomenology, research into first person accounts of experience. It does this by using distinctions in language, the internal sensory representations, and imagery. It draws on the NLP concept that people meld sights, sounds, and feelings before they speak (again, we reiterate the point in this thesis that NLP refers to Neuro-Linguistic Programming and not the commonly referred term Natural Language Programming). With the endorsement of NLP's founder, Dr. Richard Bandler, it has grown into an area called Medical Neuro-Linguistic Programming (Thomson, 2015) with a focus on the use and meaning within language to improve health.

The Preferred Representational System (PRS) has been used in teaching. A study of 283 teachers wanted to know if there was any modality dominance across the Visual and Auditory Representational System modalities (Tardif, Doudin & Meylan, 2015). They highlighted there is a distinction between visual and auditory modalities used by pupils in schools. In another study, testing student's preferences across the visual, auditory, and kinaesthetic Representational System modalities and helped teachers prepare lectures better (Ancusa, Bogdan & Caus, 2013).

However, Gray (2012) draws on neuroscience to understand the tenets that underpin NLP and suggests it is underpinned by neuroscience, and that our perceptions are reshaped by memory, expectation, cognitive filtering and past experience and broken into a world of things and categories and the borders between objects and categories

by the brain. By drawing on Canonical neuroscience, Gray (2012) shows that NLP can integrate new learnings using NLP in less than 24 hours, and not the usual thirty days it takes to transfer long term memory from the hippocampal stores to permanent cortical networks. He highlights NLP activates a behavioural off-switch in one of the brain's known circuits consisting of the ventro-medial prefrontal cortex, the anterior and posterior cingulate giri, medial temporal lobe and the precuneus related to related to evaluation, self-control, memory, prediction of future behaviour and empathic understanding of others. If this is true then research into neuroscience and the sensory modalities might open up a new avenue to explore.

Churchland (2002), suggests the self is identifiable with a set of representational capacities of the physical brain, drawing on the 18th-century philosopher David Hume's description of self to highlight it as a collection of changing visual perceptions, sounds, smells, feelings, emotions, memories, and thoughts, etc. While perception has been viewed as a modular sensory modality function, Shimojo and Shams (2001) and Yan *et al.,* (2017), suggest they are not separate modalities. They suggest that a unified consciousness, another word for self, is constructed from cross-modal inputs (Winkielman, Ziembowicz & Nowak, 2015).

The U.S. Institute of Medicine of the National Academy of Science was commissioned in the late 80's to investigate neuroscience techniques (Martin & Pechura, 1991). Of specific interest was functional Magnetic Resonance Imaging (fMRI). Many studies have been conducted in to the area of the sensory modalities since, using fMRI, for example; visual and auditory areas (Linden *et al.*, 1999), gustatory and somatosensory perceptions (Cerf-Ducastel *et al.*, 2001), olfactory (Gottfried *et al.*, 2002), visual and kinesthetic (Gulliot, 2009), and more interesting, haptic tactile imagery (Yoo *et al.*, 2003; Deshpande *et al.*, 2008). The term haptic, a bidirectional sensory modality includes an awareness of the outer surface of the body (tactile), and movement, muscle tension and limb position (kinesthetic) (Tan, 2000). This is a wider definition than kinaesthetic that was used in the earlier NLP studies.

What becomes clear is that there are cross-modal binding and integration of each modality (Calvert, Campbell & Brammer, 2000; Shimojo & Shams, 2001; Driver & Noesselt, 2008; Blank, Kiebel & von Kriegstein, 2015; Brunel, Carvalho & Goldstone, 2015). This is an important concept because the results of the earlier sensory study (Kernot, 2013) looked at each word as a single modality, and by considering the cross-

modal aspects, the unified consciousness, or self can be represented through a more refined sensory algorithm that spans several modalities.

2.4.1 Key Cross-Modal Studies

In a study of 523 concrete object nouns by 420 undergraduate students, Amsel, Urbach, and Kutas, (2012) categorized each noun on the five sensory modalities, colour (Visual), sound (Auditory), graspability (Haptic), smell (Olfactory), taste (Gustatory), and on the motor modalities, motion and pain. According to Amsel *et al.* (2012:1030):

> "*Each of the five traditional Aristotelian sensory modalities (vision, touch, hearing, smell, and taste) is represented," (*in this study*) "in addition to the sensation of pain. We assessed two kinds of visual knowledge, colour, and motion, which are represented in different brain regions proximal to the corresponding sensory cortex.*"

Each noun object was scored across the seven categories and given a numeric score from 1 to 8. They also provided a value for word familiarity, but no single dominant modality was provided and highlighted the value of concreteness as a key term in assessing sensory words.

In a different study of 423 sensory-based prenominal adjectives by 55 native English speakers, Lynott and Connell (2009), collected words from a range of sources and categorized each object on the five sensory modalities. Lynott and Connell (2009:560) found a 74.8% variance from two factors principal component analysis. Their analysis highlighted significant correlations for the majority of modality pairs, although auditory ratings correlated negatively with all the other clusters and suggested that auditory experience has little to do with other types of perceptual experience. The strongest positive relationship was between olfactory and gustatory modalities, and to a lesser extent, a positive relationship also appeared in the visual-haptic cluster. Only gustatory and haptic ratings showed no appreciable relationship.

What is important from Lynott and Connell's (2009:526) study, is that they concluded most sensory-based words are multimodal rather than unimodal with clustering in the visual-haptic and olfactory-gustatory modalities.

In another study, this one of 400 nouns by 34 native English speakers, Lynott and Connell (2013), obtained nouns from the MRC psycholinguistic database (Coltheart, 1981; Wilson, 1987) to generate a random list. They categorized each noun object on the five sensory modalities. Each noun was scored as a percentage of the mean, as was

word familiarity. What was different from the Amsel *et al.* (2012) study was that Lynott and Connell provided a dominant modality and exclusivity percentage. Drawing on an earlier study about adjectives, Lynott and Connell (2009) discovered that concepts using nouns are more multimodal across the range of the five sensory modalities than adjectives. This suggested that prenominal adjectives, words that immediately precede the noun, appear in fewer of the five sensory modalities.

While the approach of both studies by Amsel *et al.* (2012), and Lynott and Connell (2013) use excellent sources of research data to categorize nouns by their modalities, any content analysis of text will be highly reliant on the occurrence of those nouns. Because Lynott and Connell (2013) highlight the benefit of prenominal adjectives over nouns, that there is more of a likelihood to have a more dominant modality, and that because a smaller set of adverbs, particularly prenominal adjectives, will occur more often over a wide range of nouns, sensory-based adverbs would seem a better approach to use for content analysis.

van Dantzig, Cowell, Zeelenberg, and Pecher (2011), drew on the results of the Lynott and Connell (2009) study. They collected modality ratings for a set of 387 properties, each paired with two different contexts to create 774 concept-property items rated through five perceptual modalities. They computed the degree a property is perceived exclusively through one sensory modality and provided modality exclusivity scores for the 387 words to a higher level of fidelity than previous studies.

2.4.2 Summary

In this section, we have discussed the recent developments in the literature in NLP and neuroscience within the context of the sensory modalities studies using NLP predicates. By using the van Dantzig *et al.* (2011) data, better fidelity sensory scores that reflect a person's use of sensory words should be realised. The concept of self through the multi-modal exchange of sensory information within the brain should also provide a better indication of authorship identification.

2.5 Identification of Self

Few authorship identification techniques attempt to identify self from the way a person thinks. But linguistic characteristics can be drawn from a person's writing style, and traces of their personality extracted to assist in their identification (for example Iqbal *et al.*, 2013; Argamon *et al.*, 2003; Argamon *et al.*, 2007; Argamon *et al.*, 2009; Zheng *et al.*,

2006; Northoff *et al*., 2006). By drawing on the cross-modal aspects of the brain, the unified consciousness, or self can be represented through the sensory modalities. It can be further strengthened by incorporating personal pronouns that describe a person on a continuum and draws from a person's gender aspects. And it can be further strengthened through a person's cognitive state by drawing on particles to characterise a person through their cognitive state. By using word features that tap into the way a person thinks and to characterise a person's writing quantitatively through the combination of referential activity, richness, gender, and sensory adjective scores, it should be possible to construct a multi-dimensional stylistic signature of a person that reflects self.

While Bucholtz and Hall (2005) suggest that identity is the product rather than the source of linguistic and other semiotic practices and therefore is a social and cultural rather than primarily internal psychological phenomenon, we prefer to take it as an internal psychological phenomenon that blends the body and mind and call self. We highlight this, drawing on Daly et al.'s(2018) links between body, personality and identity, where they note that while all illnesses can alter physical abilities and change relationships, it is the brain's neurologic disorders such as Alzheimer's disease and Parkinson's disease that can uniquely alter fundamental personality traits that contribute to identity. To help define identity in the context of this thesis, we center our idea of identity in the concept of embodied cognition. Embodied cognition is grounded in cognitive neuroscience and psychology, and research into it has risen exponentially over the last 25 years (Gjelsvik, Lovric, & Williams, 2018). Embodied cognition regards the human body and the environment as significant factors in the way we think and feel (Guell, Gabrieli, & Schmahmann, 2018). This is done by processing both emotional and modality-specific systems in the brain (Barsalou et al, 2003; Niedenthal et al, 2005; Mahon, 2015; Tillman, & Louwerse, 2018), where both emotion and the sensory multi-modal specific processing of memory work together (Niedenthal, 2007; Dreyer & Pulvermüller, 2018), also known as semantic cognition (Ralph et al, 2017:2). Embedded cognition is grounded in the idea that the body is critical in idea generation and then acting on those ideas, or thoughts. When objects and events are viewed through the eyes of the self they typically become emotionally coloured, and thereby more intimately related to one's sense of self (Northoff et al., 2006: 441).

2.6 Extremist's Disorders

Extremists and fundamentalists appear throughout history and exploit different religious and ideological beliefs to justify their violent behaviour (Simpson, 2014). Stottlemyre (2014) describes one aspect, lone wolf domestic terrorism as something that has been around since at least the nineteenth century. He references recent studies (Spaaij, 2012), that state the differences between lone wolf and group-affiliated terrorists are lone wolfs' motivations tend to be based on personal grievances, and political or religious ideologies, while the terrorist group's motives are almost exclusively ideologically based. This view of the lone wolf is reflected in the view of the 'green-on-blue' insurgent (Bordin, 2011). Lone wolf terrorism is often considered more dangerous than attacks conducted by terrorist groups. Because loners do not need to interact with or receive funding from a group, it is very difficult to track their movements or even their existence. It is difficult for law enforcement officials to detect when an individual becomes radicalized, whereas the ideology and general intent of terrorist groups are often widely publicized (Bakker & DeGraaf, 2010). For all these reasons, lone wolf terrorists often do not end up under scrutiny from law enforcement officials until they have already conducted at least one attack says Stottlemyre (2014).

Having discussed how to identify a person through self across features such as gender, Referential Activity, Richness, and the sensory aspects of adjectives, and highlighting the value of critical slowing down as a way to identify a tipping point, it is important to make a few brief points about research that indicates the mental state of extremists such as terrorists and lone wolves. Of note, Canadian lone wolf, Martin Couture-Rouleau, who killed a soldier outside Montreal, appeared to be depressed in a similar way to Justin Bourque, who shot and killed three police officers in Moncton, New Brunswick earlier the same year (Simpson, 2014). There are links to depression, suicide and mental illnesses in the violent extremist. Mass murderer's lives are plagued with psychosis, paranoia, depression, while lone wolves typically suffer from mental illness and tend to be suicidal (Capellan, 2015: p4). With suicide terrorists, mental health problems, personal crises, coercion, fear of an approaching enemy, or hidden self-destructive urges play a major role (Lankford, 2014). Bobadilla (2014) suggests these self-destructive urges might be from *vulnerable* narcissism, and these timid and shy characteristics that hide the narcissism are related to avoidant personality disorder (see Meyer, Ajchenbrenner & Bowles (2005) and the comments about sensory sensitivity and high levels of depression in APD patients) that has been observed in would-be

suicide bombers. As Lankford (2014: 351), so eloquently puts it: "By better understanding suicide terrorists, experts in the behavioural and brain sciences may be able to pioneer exciting new breakthroughs in security countermeasures and suicide prevention."

It is known that lone wolf actors broadcast their intent, also known as seepage or signaling. While lone wolves may isolated from the physical world, they still communicate by the following means: threatening statements, letters, manifestos, and videotaped proclamations on the Internet that refer to a future attack (Hamm & Spaaij, 2017). Most lone actors are generally poor at maintaining operational security, and they leak their motivations and capabilities months or years before an attack (Schuurman et al., 2018), so that other people generally know about the offender's grievance, extremist ideology, and intent to engage in violence (Gill, Horgan, & Deckert, 2014).

2.7 Classification Techniques

There are many different techniques that can be used for authorship identification. Many involve counting the frequency of word types, looking at the length of sentences, or identifying commonly used keywords, or n-grams (Craig & Kinney, 2009; Vickers, 2011, 2014). Rudman (1998) identified over 1000 different possible features that could be used, but that a serious problem exists. Rudman (2012) reviewed an additional 600 studies with the same conclusion.

In their review of stylometric writing techniques, Zheng *et al.* (2006), categorised authorship analysis studies into three major fields. They defined *authorship identification,* or *author attribution* as a technique to determine the likelihood that a piece of writing is written by a particular author, by examining other works by that author. They defined *authorship characterisation* as a technique that summarised the characteristics of an author, such as gender, and cultural background to create a profile of the author. This second technique does not draw conclusions from the works of other works of known authors, but can be used to identify similar characteristics of an unknown author. This technique was used earlier by Kernot (2013). Zheng *et al.*'s (2006), third technique is *similarity detection*, which compares multiple pieces of writing, to determine if a piece of writing was produced by a single author. All three techniques have their place in determining authorship, but the focus of the program of research reported in this thesis is in characterising unknown authors.

After reviewing the techniques used in 23 previous studies across 270 lexical features, Zheng *et al.* (2006) tested four types of features across three popular, yet powerful, classification techniques: Support Vector Machines (SVM); decision trees; and feed-forward networks, and determined that the SVM technique outperformed the others. However, over the past few years, big data and deep learning is becoming a very popular methodology, while the size of the dataset is relatively unimportant because samples are taken and processed using small sets of contiguous words to create penta-grams and small image sets with few deep learning layers (Gaonkara *et al.,* 2016; Hassan *et al.,* 2017). Making sense of big text to visualise spatial and temporal relationships data can be achieved through tools like Leximancer and Discursis where each speaker's text segments can be visualised in a sequential way, gripping up each of their texts temporally, or using concept mapping to generate themes that can be visualised spatially across different authors (Angus, Rintel & Wiles, 2013; Stockwell et al., 2009).

Stamatatos (2009), also surveyed modern techniques and emphasised Rudman's (1998) disappointment as to the state of authorship attribution after 300 publications over a period of 30 years, concluded that variations in text length, number of authors, and amount of training texts had an effect in determining the accuracy of results. Notwithstanding this, authorship attribution from text is admissible in court in some US jurisdictions because of the contributions from Abbasi and Chen (2008), Benjamin et al. (2014), and Chaski (2005; 2001).

In reviewing Samuel Taylor Coleridge's 1816 work (edition of Coleridge 1984), Benatti and Tonra (2015) selected a supervised and an unsupervised approach to authorship identification. They experimented with two unsupervised methods (Cluster Analysis and Principal Component Analysis) and three supervised ones (Support Vector Machines, Nearest Shrunken Centroids and k-Nearest Neighbours), but in their paper, they limit their discussion to Cluster Analysis and Support Vector Machines, and highlight that neither the unsupervised nor supervised methods provided a probable attribution of authors.

Stylometric analysis has been extensively used to determine the authorship, from the undocumented collaborations of the playwrights in the Elizabethan period (Segarra *et al.*, 2017) to recent Prime Ministers (Garrard, 2009; Snowden, Griffiths & Neary, 1994) and famous novelists (Garrard *et al.*, 2005; Le *et al.*, 2011).However, there is dissension among leading scholars about an agreed method (Rudman, 1998; 2012; 2016;

Stamatatos, 2009), but the most successful and robust methods are based on low-level information such as character n-grams (a contiguous sequence of n items of text) or auxiliary words (function word, stop words such as articles and prepositions) frequencies (Stamatatos, 2009). The premier work in evaluating authorship includes MacDonald P. Jackson, Brian Vickers, and Hugh Craig and Arthur Kinney (Segarra *et al.*, 2017). Jackson (2006) uses common low-frequency word phrases, repetition of phrases, collocation, and images to link word groups to other works. Vickers (2011) uses a tri-gram, or n-gram, approach instances where three consecutive words in a sentence closely match known authored works, while Hirch and Craig (2014) use function word frequency and other methods, that includes ones based on word probabilities and the Information Theoretic measure Jensen-Shannon divergence (JSD) and unsupervised graph partitioning clustering algorithms (Arefin *et al.*, 2015). For a detailed explanation of these techniques, we recommend Juola (2008). The meaning-extracting method (MEM) from the field of psychology to extract themes from commonly used adjectives and describe a person from their personality, or self is very different (Boyd & Pennebaker, 2015; Chung & Pennebaker, 2008). The methodology employed within this research thesis also focuses on personality as a driver and looks at *why* people say the things they do, their particular word choices, and how those aspects can create a unique stylistic fingerprint of a person.

In discussing the different methods of authorship attribution, Juola (2008) concludes that the best choice of the feature set is strongly dependent upon the data to be analysed, and no method has yet emerged from any study as being particularly good within a narrow range of language, genre, size, etc. Rudman (2012) revisited the problem, 13 years after his earlier critique, after well over a further 600 studies and concluded that there is still no consensus as to correct methodology or technique. Perhaps the state of continued flux in the determination of a small group of generic methods is unachievable because of the complexity of the problem. Rudman's view seems as consistent today as it was in Rudman (1998: 360): "One of the most important facts to keep in mind is that each authorship study is different. Not only are there the various types but each author, each genre, each language, each time period force variations on the experimental design and require unique expertise."

However, five key points emerge from Juola's (2008: 319-324) recommendations that are critical in the determination of methods for this study. They are:

- Methods using a large number of features seem to outperform methods using a small number of features, provided that there is some method of weighting or sorting through the feature set.

- Methods that do not use syntax in one form or another, either through the use of word n-grams or explicit syntactic coding tend to perform poorly.

- Simple unsupervised analysis — most notably, principal component analysis (PCA) — will sometimes produce useful and easy-to-understand results. On the other hand, PCA and similar algorithms are often unable to uncover authorship structure that more powerful algorithms find.

- The same vector space that categorizes text can be used to categorise individual words (or features); one can literally superimpose on the graphic separating A from B the words used, and the words near A are the ones that A uses and B does not.

- The real heavyweights emerging from Juola's Ad-hoc Authorship Attribution Competition (AAAC) are the same high-performing analysis methods that have been used elsewhere. These high-flyers include SVMs, linear discriminant analysis (LDA), and k-nearest neighbour in a suitably chosen space.

2.8 Philosophical Theory

This research thesis sits within the positivism area of philosophy. Epistemologically, we focus on discovering observable linguistic measures from a wide range of writing, including plays, poems, novels, suicide notes, manifestos, and a range of internet blog posts and newspaper articles. We use quantitative methods of analysis to develop hypotheses, which we test through a series of experiments. Using a positivist approach and experimental design with meaningful data, we conduct experiments and test hypotheses. We use a mathematical modelling approach to personality or self.

In this research thesis, we use a multi-disciplinary approach and examine a person's identity using personality through a complex systems lens. We draw on a number of disciplines, including computer science, ecology, linguistics, mathematics, neuroscience, psychology, and statistics. The overarching scientific discipline is neurolinguistics, a scientific discipline that studies the relationships between the human brain and language. While it has been around in one form since the time of the Ancient Egyptians (Gross, 1987), in its current form, it is relatively new (Leikin, 2016),

and recent neurolinguistics studies have started to investigate the personality aspects of a human language to improve authorship profiling (see Pennebaker *et al.*, 2015a; Litvinova *et al.*, 2016; Skillicorn *et al.*, 2017). The ability to profile user personality, particularly by inferring stable differences in individual behaviour from writing, can be used to predict a person's preferences and future behaviour with sufficient accuracy (Wright & Chin, 2014).

2.9 Summary

In this chapter, we have built on earlier research on gender pronouns and sensory-based predicates and identified new key linguistic features from ecology, neuroscience, and neuropsychology and called it RPAS, where the Sensory (S) comprises of a multimodal classification of the five senses (VAHOG) to identify a person. In the next chapter, data from the van Dantzig *et al.* (2011) study will be defined to construct sensory-based adjective references for scoring an author. Data from the MRC Psycholinguistic database will be defined to construct referential activity predicate references. The existing gender pronoun data from Kernot (2013) will be used to focus on the gender score as well as the Masculine / Feminine classification. We will discuss the research methodology that is used, and we focus on four specific aspects that have come about from this literature review to extract features that reflect self through RPAS. We draw on the idea of Critical Slowing Down, and the recommendations highlighted above from Juola (2008) to answer the research question.

Methodology

In this chapter, we address the Data and Algorithms Development phase. We describe the approach undertaken to create the author signatures so that each individual author's work can then be tested against a series of experiments. We describe how the reference data list is created, and how the works are reduced into a Bag of Words (BOW). We identify the Richness (R), Personal Pronouns (P), Referential Activity Power (A), and Sensory-based adjective (S) equations that will be used to create a signature of an author that describe self. We describe how the research hypotheses are tested through a series of studies.

3.1 The Approach

Given that the general consensus (as discussed in 2.7 Classification Techniques) is that methods need to change to suit the data and the problem, the exact methods used vary with each study or experiment.

However, we begin with an overview, and some aspects of the method will remain consistent. The data, once identified, will be tokenised into a Bag of Words using the Stanford Parts of Speech tagger[6] (Toutanova & Manning, 2000; Manning, 2011), and converted to lowercase. The Stanford Parts of Speech tagger was chosen over other taggers because it uses the very popular English language Penn—Treebank phrase structure tags, is open source, but has other language options. Some variation is likely to occur in the tagging of speech types, however this is unavoidable because there is no one single tagger. All punctuation and symbols including numbers will be removed before the data is aggregated into word lists. The identified features that map to RPAS will be processed using the reference lists identified below, and each word category will be further processed using the equations below, to create a signature of each author's work that reflects self.

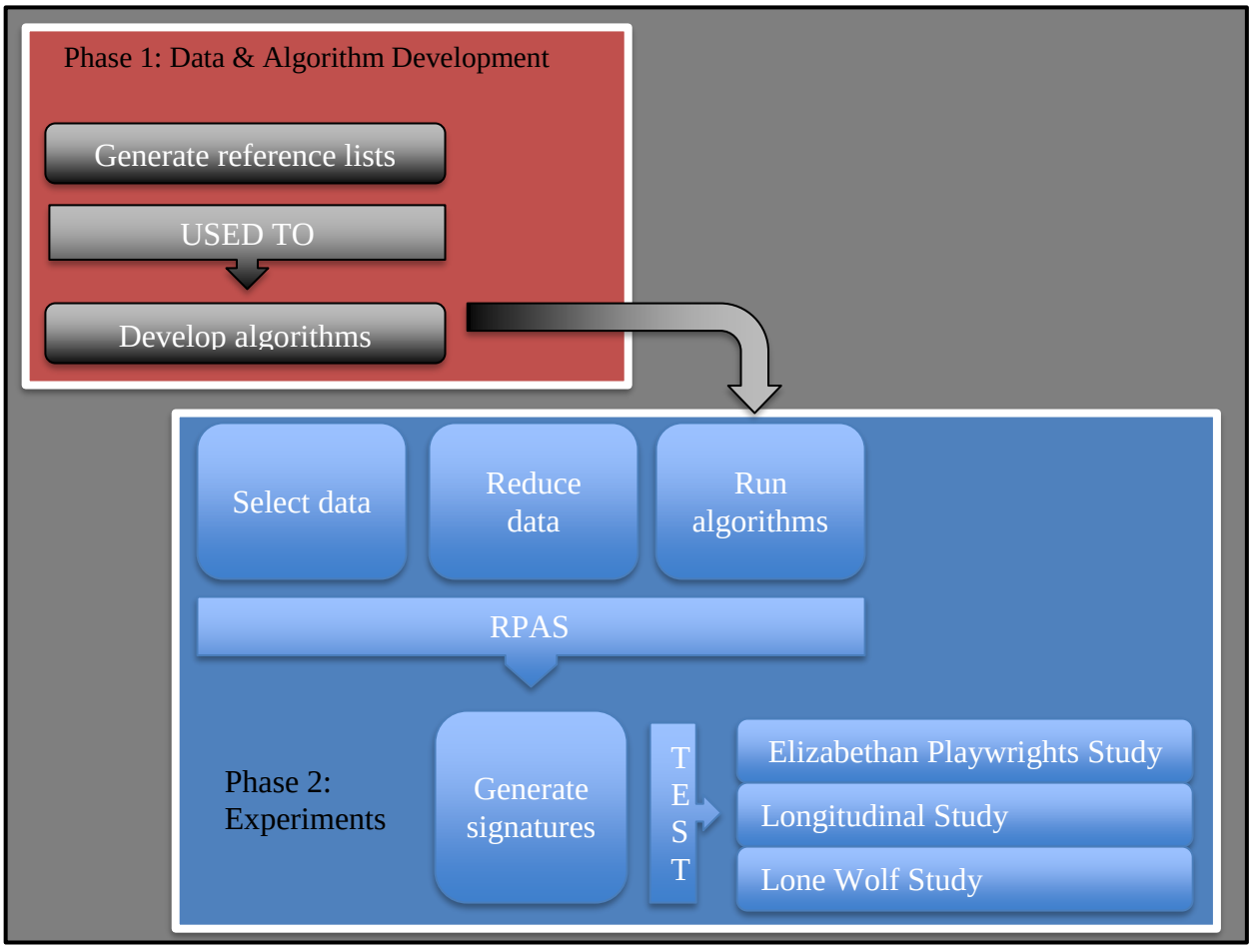
As highlighted in Juola (2008), a vector space model that categorizes text is of value, as is SVM, linear discriminant analysis (LDA), and k-nearest neighbour. Therefore, where

[6] The Stanford Part-Of-Speech Tagger is available from the Stanford Natural Language Processing Group, Stanford University. Available at: http://nlp.stanford.edu/software/tagger.shtml

possible a combination of one or more techniques will be used to test the author's signatures for each experiment. We will draw on the recommended clustering approach in Burns and Burns (2012), and conduct Two Stage Hierarechical (Agglomerative) Clustering (HAC) using Ward's Method and Squared Euclidean Distance. We will use a vector space model that maps cosine and minmax similarity detection techniques. We will use the intruder method (Koppel & Winter, 2014) and use k-nearest neighbour techniques through the R package seriation (Buchta, Hornik & Hahsler, 2008) that appear in some of the seriation methods (TSP- Travelling Salesperson, Chen – Rank-two ellipse seriation, ARSA – Minimise Anti-Robinson events using simulated annealing, HC – Hierarchical clustering, and OLO – Hierarchical clustering with optimal leaf ordering).

Figure 3 highlights the general method to be applied, and Table 2 highlights the aims and the hypothesis to be tested for each study. Table 3 then summarises the techniques used prior to a detailed description of the equations and techniques.

Figure 3: Methodology



3.2 Research Hypothesis Experiments

In this section, we link the research hypotheses to the series of experiments. There are three planned parts of experiments designed to answer the research question by testing the hypotheses. They are, Chapter 4 – 6, the authorship identification of the Elizabethan playwrights; Chapters 7-9, changes in an author's signature over time, using Shakespeare, Iris Murdoch and P.D. James; and Chapter 10, lone wolf signature classification. Across the three parts, seven studies will be conducted to test the four hypotheses (stated below in Table 2) and answer the research question.

Table 2: Chapter Experiments, Aims, and hypotheses

Chapter	Study	Aim	Test Hypothesis
4 – Shakespeare, Marlowe, and Cary	1	Identifying features to identify authors (known authorship) from RPAS	H1
5 – Edward III	2	Testing RPAS on an unknown author.	H1
6 – Passionate Pilgrim	3	Testing RPAS on a set of multiple unknown authors.	H1
7 –The Sonnets	4	Identifying features of a single author to identify characteristics in small texts that change over time from RPAS.	H2
8 – Iris Murdoch and P.D. James	5	Identifying features of two authors to identify characteristics that change over time from RPAS when one has proximal events and life stressors that mimic a potential terrorist.	H2
9 – Iris Murdoch and P.D. James	6	Identifying characteristics that change over time using tipping point techniques that visualises the Critical Slowing Down phenomena from RPAS prior to a life changing event.	H3
10 – Lone Wolf	7	Identifying different RPAS features of lone wolf suicide attackers that are different from normal or depressed writers.	H4

In Table 2, the texts were chosen specifically to start off testing some simple tests and have them build in complexity. Shakespeare and Marlowe were chosen because of the well-researched author identity claims that could be validated. We include Elizabeth Cary to add an independent element of gender to the study. The Edward III and the Passionate Pilgrim also add a level of increasing complexity to the testing using multiple known and unknown authors. The Sonnets were chosen to transition the approach from the Elizabethan period when looking at change in writing to

contemporary writing. Murdoch and James have been well researched as longitudinal studies of linguistic change in individuals that have died from normal aging and mental disease. Lone wolves are a current problem that frames the research clearly in the national security space which is fitting for the national security college but also it is underpinned by the concepts from the previous studies.

Table 3: A summary of all the techniques employed

Study	**Technique**
Shakespeare's Sotto Voce	
1	POS Analysis of Shakespeare data
1	Pearson's Product Moment Correlation, r, analysis
1	PtoR, AtoR, and StoR plots
1	Word Accumulation Curves (Richness testing alternative)
1	Hierarchical Clustering Analysis (HCA)
1	Principal Component Analysis (PCA)
1	Stepwise Linear Discriminant Analysis (LDA)
1	Partial synthetic data approach to testing results
Kyd and Shakespeare's Edward III	
2	PtoR, AtoR, and StoR plots
2	Vector Space Method (VSM)
2	Imposters Method
2	Seriation – with noise
Study	**Technique**
Shakespeare's Passionate Pilgrim	
3	PtoR, AtoR, and StoR plots
3	Principal Component Analysis (PCA)
3	Linear Discriminant Analysis (LDA)
3	Vector Space Method (VSM)
3	Seriation – with noise
Shakespeare's Dark Lady	
4	Seriation – with noise
Murdoch and James – Depression and Apathy`	
5	POS Analysis of Murdoch and James
5	Mann-Whitney U tests on Richness
5	Content to Function Word ratios
5	Mann-Whitney U-Tests on Sensory words (S)
5	Principal Component Analysis (PCA) on VAHOG
5	Content word comparisons in different time periods (10-12 years)
5	Function word comparisons in different time periods (10-12 years)
Murdoch and James - Impacts of aging and life events on identity.	
6	Richness (R) comparison of Murdoch and James.
6	Personal Pronouns (P) comparison of Murdoch and James.
6	Referential Activity power (A) comparison of Murdoch and James.
6	Sensory Adjectives (S) comparison of Murdoch and James
6	CSD - 1-lag autocorrelation (AR1) on Sensory Adjectives (S)
	Continued next page

Study	Technique
Murdoch and James - Impacts of aging and life events on identity (cont.)	
6	CSD - Fisher-Pearson coefficient of skewness (G1) Sensory Adjectives (S)
6	Principal Component Analysis (PCA)
6	Stepwise Linear Discriminant Analysis (LDA)
Suicide Attackers	
7	LIWC sentiment analysis Mann-Whitney U test to compare normal to Attackers
7	Stepwise Multiple Regression Analysis on RPAV
7	5-fold cross-validation
7	Mann-Whitney U tests on 7 sentiment tags
7	Comparison of Depression and terror suicidality

3.3 Reference Lists

Drawing on the work on the existing work on gender, and the new work on the Referential Activity and sensory adjectives, three reference lists underpin the equations in Section 3.4. The reference lists for each of the three feature sets are included in the Appendix A. For a brief explanation, a summary of the concepts is described below:

3.3.1 Gender

The gender reference list began from the Argamon *et al.* (2003) study. In it, the different pronoun types were listed by their assumed gender and are labelled M (More likely used by a male gender) or F (More likely used by a female gender) based on the statistical results. From the results of earlier gender testing, 27 pronouns were identified (Kernot, 2016) and highlighted the results of the accuracy by words from the earlier experiment.

3.3.2 Referential Activity Power.

The Medical Research Council (MRC) Psycholinguistic Database is a computerized database of psycholinguistic information with 98,538 words (Colheart, 1981) and it has a number of key linguistic word properties, including scores for imageability and concreteness. Only particles (articles, pronouns, conjunctives, and prepositions), words that reflect self, were extracted from the MRC Psycholinguistic database. Within the database, each word had a score out of 700 for both imageability and concreteness. The higher the score, the more the word held imagery and concrete aspects from Referential Activity. All words that scored zero in either category were removed, and this left a word list of 117 highly concrete and imagery words. The scores

from each word category were divided by 700 to convert them to a percentage. Both categories were then added together and averaged. For example, only a word that scored 700 in both imageability and concreteness would obtain a score of 1. In the case of the words in the list, they ranged from 0.27 to 0.85. This list appears in Table 33 of Appendix A. Referential Activity examples would include: Articles – 'a', 'an', 'the'; Prepositions – 'about', 'after', 'to'; Conjunctives – 'although', 'else', 'once'; Pronouns – 'all', 'every', and 'our'.

3.3.3 Sensory-based Adjectives

This list drew on the work of van Dantzig *et al.* (2011), and their collected modality ratings for a set of 387 properties, each paired with two different concepts to create 774 concept-property items and rated through five perceptual modalities. They computed the degree a property is perceived exclusively through one sensory modality and provided modality exclusivity scores for the 387 words. In the list, each word (sensory adjective) has two entries. If the word is perceived exclusively in one sensory modality, then both entries will be the same modality. If this is not the case, then the word is bi-modal, and each of the two highlighted modalities assigned to it will be different. The words in the list ranged from 0.1 to 0.98. This list appears in Table 34 of Appendix A. Sensory Adjective examples would include: Visual – 'abrasive', 'big', 'immense'; Auditory – 'banging', 'barking', 'plain'; Haptic – 'abrasive', 'immense', 'lukewarm'; Olfactory – 'aromatic', 'garlicky', 'lemony'; Gustatory – 'creamy', 'garlicky', and 'tender'.

3.4 Equations

In this section, we identify the equations used to measure the RPAS and the CSD dynamical property, and they are described in detail below.

3.4.1 The RPAS method

In this section, Richness (R), Personal Pronouns, or internal gender (P), Referential Activity Power (A), and the Sensory-based Adjectives (S), also known as RPAS are described in detail. We employ a new methodology that adopts a multi-faceted approach to text analysis and reveals details about a person's personality; their sense of self, from subtle characteristics hidden in their writing style (Argamon *et al.*, 2009; Iqbal *et al.*, 2013; Northoff *et al.*, 2006). **RPAS** draws on biomarkers for creativity and known psychosis (Rosenstein *et al.*, 2015; Zabelina *et al.*, 2015) to identify characteristics within

an author's writing. **RPAS** comprises: **Richness (R)** (Menhinick, 1964; Tweedie, & Baayen, 1998), the number of unique words used by an author is linked to a person's education level and also their age, where it can grow up to about 65 years of age (Hartshorne & Germine, 2015); **Personal Pronouns (P)** (Argamon *et al.*, 2003; Kernot, 2016; Pennebaker, 2011; Pennebaker, Mehl, & Niederhoffer, 2003), the pronouns used, closely aligned to gender and self; **Referential Activity Power (A)** (Bucci, 2002; Bucci & Maskit, 2004), based on highly concrete and imagability function words, or word particles, the concept has been used in clinical depression studies, and we draw on the Medical Research Council Psycholinguistic Database of English words; and **Sensory (S)** (Kernot, 2013; Lynott, & Connell, 2009; Miller, 1995; van Dantzig *et al.*, 2011; Fernandino *et al.*, 2015), five sensory measures (V-Visual A-Auditory H – Haptic O – Olfactory G - Gustatory) corresponding to the use of the senses.

3.4.1.1 Richness (R)

The Richness equation is a measure of a person's ability to use a vocabulary of a determined size, and for two documents of the same length, the one with more different (unique) words – a larger vocabulary – has greater richness. While the values of lexical richness change for different measures used because of text length, it is necessary to correct for text length (Tweedie & Baayen, 1998)), we do this with ratios (Singhal, Buckley & Mitra,1996; Kessler, Numberg & Schütze, 1997) because we are effectively examining the word density within each work and comparing it to the others (Gotelli & Colwell, 2011). Any global richness coefficient can be ignored in this case. The formula is given as:

Equation 1: Richness

$$\text{Richness (R)} = \frac{w}{N}$$
$$\text{where } w = \text{number of unique words in the document, and}$$
$$N = \text{total document word count.}$$

There are theoretical limits to this equation, and the size of documents must be matched to avoid artefacts. Eventually, the value will reach an asymptote as no new words are found. Near that point, the larger the document size, the smaller the Richness score will be ($0\ as\ N \to \infty$).

The type-token ratio (TTR) is the ratio of vocabulary size, to the text size, in log form (Tanaka-Ishii & Aihara, 2015) and it can be considered a similar, but an inverted

variant of Menhinick's (1964) species diversity equation (Equation 1) that measures vocabulary richness. TTR is one of the oldest and easiest ways of measuring richness but it is dependent on text size, and while many attempts to reduce this problem have been proposed no one has been fully successful (Kubát, & Milička, 2013).

While TTR and its inverted reciprocal, the mean word frequency is affected by text length size, other alternate measures such as Herdan's C, Simpson's D, Honre's H, Janenkov and Nesitoj's LN, Yule's K, Guiraud's R, Sichel's S, Dugast's U, Herdan's V, Brunet's W, and Orlov's Z were also found to be impacted by text length, the exception was Yule's K and Orlov's Z but given their within-text variability, they should be used with care (Tweedie & Baayen, 1998).

In 1944, Yule introduced his author identification measure, Yule's K to differentiate authors, and found that it converges to a value for a certain amount of text and remains invariant for any larger size, and its value could be considered as a text characteristic (Tanaka-Ishii & Aihara, 2015). Linguist George Zipf popularised an observation made by earlier scientists, that given a corpus of natural language utterances, word frequency is inversely proportional to its rank in the frequency table, so that the most frequent word occurs approximately twice as much as the next word, and it three times as much as the next, and so on in a power law relationship (Powers, 1998). In 1983, Orlov and Chitashvili considered the long tails of vocabulary distributions and deduced a parameter, Z, for which Zipf's law does not hold, but Orlov's Z is an extension of Zipf's law and shows that the expected value of the vocabulary size for a given text can be mathematically determined using a sole parameter, Z (Kimura & Tanaka-Ishii, 2014).

Kimura and Tanaka-Ishii (2014) examined Tweedie & Baayen's (1998) report that Yule's K and Orlov's Z were convergent across a range of different languages, and found that Yule's K using both a small 170, 000 word English corpora and a large 18,000,000 corpora was convergent at 100,000, but that Orlov's Z was not convergent across the same data. In a separate report, Tanaka-Ishii and Aihara (2015) retested Yule's K using smaller length documents and found that at a document size of 10,000 or greater the index was convergent, but that smaller texts did not possess the discriminatory power of author identification for which Yule had hoped.

Wimmer and Altman (1999) also reviewed Tweedie and Baayen (1998) and highlighted there were many different technique index values and different behaviours of their

asymptotes (the point at which the number of new words introduced is no longer effectively increasing). They stressed there was a need to understand the origin of each index and its underlying behaviour.

In a study of Dutch first language and Dutch second language children using TTR, Vermeer (2000) tested it against two alternate approaches, Guiraud's R and the Uber index. The index of Guiraud (1960) is similar to TTR, but it uses the square root of N (the total number of words in the document). While Dugast's (1978) Uber Index is a log function variant of the TTR. Vermeer (2000) suggested that by keeping the acquired word count below a total of 3000 that the number of types (which are a part of TTR) and the Guiraud and Uber indexes seem to do well. They also noted the impact of a person learning function words has an impact on the different types of lexical richness measures.

Building on Vermeer's (2000) approach, Van Gijsel, Speelman, and Geeraerts (2005) suggested that an alternative for the simple TTR is the Mean Segmental TTR (MSTTR), and by using text sections of equal length, richness scores worked well from 750 up to 1350 tokens.

In a later study of 4175 tweets from 14 Twitter users' posts in English and Spanish, Juloa and Mikros (2016) showed that there is a very high correlation between ordinary stylistic variables measured on the two languages use using word length, which is often viewed as a proxy for vocabulary richness and complexity, the traditional type token ratio (TTR), hapax legomena (words that appeared only once), Yules K, and a collection of vocabulary richness measures from Kubat, Matlach, and Cech's (2014) QUITA software package. They suggest that TTR and Yule's K were different, but generally most of the correlation between the various measures was extremely high, and they appeared to be generally measuring very similar things in the data.

In summary, for text sizes greater than 10, 000 words that Yule's K, or better still, Rényi's higher-order entropy (this is a generalisation of the Shannon entropy (Shannon, 1948), an effective measure of uncertainty in the field of information theory) perform well and are independent of text size (Kimura & Tanaka-Ishii, 2014; Tanaka-Ishii & Aihara, 2015). For smaller texts the size of a tweet, or up to 3,000 words, the well-used type-token ratio (TTR) is sufficient when compared to many alternative techniques as they all suffer from increasing file size (Juloa & Mikros, 2016; Van Gijsel, Speelman & Geeraerts, 2005; Vermeer's, 2000). However, using individual files that are

identical in size, or very similar, up to 10,000 words will also reduce the error induced by the TTR (from Zipf power laws). The biggest criticism of TTR is that it should not be used on its own, rather it should be incorporated into a larger suite of techniques (Kubát & Milička, 2013; Vermeer, 2000), and we avoid this criticism by using a multivariate technique.

3.4.1.2 Personal Pronouns (P)

A person's personal pronouns use (Equation 2 or see Kernot, 2013 for further detail) provides a score that can identify an author's unique style on a continuum between 0 and 1 and can differentiate between authors of the same or different sex. The formula draws from the binary logistic regression model, also known as a logit model, where it attempts to classify or predict a discrete, categorical variable (in this case masculine M or feminine F writing) from predictor variables (here using the number of personal pronouns used in a person's writing) and it classifies it as 0 (feminine) or 1 (masculine). In this case, we draw on two existing studies on gender (Argamon *et al.*, 2003; Kernot, 2013).

The Argamon *et al*. (2003) study analysed 25 million words in 604 documents using a range of fiction and non-fiction articles (natural science, applied science, social science, world affairs, commerce, arts, belief/thought, and leisure) from the British National Corpus to assign a dominant gender across 29 statistically significant personal pronouns. These results were further distilled (Kernot, 2013; Kernot, 2016) and statistically significant gender identities determined to 90% accuracy using three personal pronouns from a collection of 25 thousand words, using articles from the internet (news reports, web articles, personal blog posts, book extracts, and an oration). The equation based on the three best predictors (the pronouns my, her, its) of a person's socially constructed gender, how they present themselves outwardly independent of their actual physical sex (Cheng *et al.*, 2011) is used to classify a person's writing.

Gender can be expressed as a Masculine (M) or Feminine (F) style. Where the Personal pronouns score is greater than or equal to 0.5, we would allocate an M categorical value, but in Kernot (2016) we also use the actual score between 0 and 1 prior to the categorical logit classification of M or F. The Personal pronouns score (Kernot, 2013) can be determined by:

Equation 2: Personal Pronouns

$$\text{personal pronouns (P)} = \frac{\exp(-0.93 - 451.86\alpha + 322.47\beta + 129.83\gamma)}{1 + \exp(-0.93 - 451.86\alpha + 322.47\beta + 129.83\gamma)}$$

$$\text{where masculine style (P)} \geq 0.5, \text{feminine style (P)} < 0.5;$$

$$\alpha = \text{'My'}, \beta = \text{'Her'}, and\ \gamma = \text{'Its'}.$$

Within this thesis, there are a number of references to the gender aspects of Equation 2, where G is either M or F referring to either Masculine (M) or Feminine (F) internal writing style. The Personal Pronouns (P) score is a number between 0-1.

3.4.1.3 Referential Activity Power (A)

Grounded in 'Critical Realism', an approach that asserts the independence of an external world whilst accepting that knowledge of that world is socially constructed and transient (Bosward, 2017), the American philosopher, Roy Wood Sellers (1959), provided a linguistic framework guided by the brain's sensory referential sensations and that concept was picked up for clinical studies into depression (Bucci & Freedman, 1981; Bucci, 1982; Bucci, 1984; Bucci & Millar, 1993; Bucci, Maskit & Murphy, 2015; Murphy, Maskit & Bucci, 2015).

Clinical psychologists use the psycholinguistic variable, Referential Activity (RA) to score a person's level of depression from their speech across the following four categories: properties of actual things or events or to anything that is experienced as a sensation or feeling sensory characteristics of language (Concreteness); the vividness and effectiveness of language in reflecting and capturing imagery or emotional experience, in any sense modality (Imagery); the quality of detail, e.g. degrees of articulation (Specificity); and, the organisation and focus (Clarity) (Bucci, & Kabasakalian-McKay, 2004; Murphy, Maskit & Bucci, 2015). While the RA measure assesses the degree to which a speaker or writer is able to translate experiences into words, it can map a continuum of a cognitive state from a healthy individual to one who has is diagnosed with depression (Bucci & Freedman, 1981).

Pennebaker *et al.* (2003) suggest that Referential Activity can be measured by a person's use of a group of function words known as particles, and include pronouns, articles, prepositions, conjunctives, and auxiliary verbs, and they can also serve as markers of emotional state and social identity.

We focus on the sensory aspects of Bucci's concepts of Referential Activity and use two of the four categories; the sensory characteristics of language (Concreteness) and the effectiveness of language to capture imagery and emotional experience in any sensory modality (Imagery). We also draw on Pennebaker *et al.*'s (2003) idea that particles can reflect the sense of a person's self, and using the Medical Research Council (MRC) Psycholinguistic Database (Coltheart, 1981), we select the particles (articles, conjunctives, prepositions and pronouns) that have concreteness and imageability scores greater than zero.

These 117 highly concrete and imageability function word scores from the MRC Psycholinguistic Database (Coltheart, 1981) were averaged for each word and these scores, ε_i can be found in the RA column of Table 33 of Appendix A. From above, we create four referential categories, one each for articles, conjunctives, prepositions, and pronouns.

For a given document, we let the number of words in each referential category, i, be ω_i and ε_i, the weight for each category then the RA Power score, A_k (Equation 3) can be determined by:

Equation 3: Referential Activity Power

$$A_k = \sum_{i=1}^{N_k} \frac{\omega_i^2 \varepsilon_i}{D}$$

where k = 1-4, N_k = 117, and D is the number of words in the document.

In the process of calculating the four elements of RA Power (A, C, P, PRON), the data is normalized based on the document or chunk size so that the ratio of Richness to Referential Activity Power becomes independent of document size. This normalised value is multiplied by its particular word weight (ε_i) and then squared by the word count frequency to emphasize the variance in the data. The four different RA Power elements are then summed to provide an overall score, (A).

3.4.1.4 Sensory Adjectives (S)

Many Sensory (S) words are processed by the brain as sight/feel and smell/taste word categories (Lynott, & Connell, 2009 For more information see Miller, 1995; Kernot, 2013; Fernandino *et al.*, 2015). We use adjectives over verbs or nouns because they appear more frequently in text and their context is not necessary. We draw on a study of 387 adjectives (van Dantzig *et al.*, 2011) that have been analysed in two different

contexts to assess the dominant visual (V), auditory (A), haptic (H), olfactory (O), or gustatory (G) sensory modality the word responds to. The study provides a list of 774 words because they were each tested in the two most dominant modalities. These 774 sensory words are allocated an exclusivity score, (φ_i) and can be found in the Exclusivity column of Table 34 of Appendix A that reflects the brain's Representational System. In this thesis, we test the concept that these values capture the sensory gating biomarker characteristics of a person which in turn can be used to construct a unique signature of a person's sensory cortex functions.

There are five sensory categories, one each for V, A, H, O, G. If we let the number of words in each sensory category, i, be w_i and ϑ_i, the weight, or exclusivity score for each category then the Sensory Adjectives, S_k (Equation 4) can be determined by:

Equation 4: Sensory Adjectives

$$S_{k\ 1\text{-}5} = \sum_{i=1}^{N_k} \frac{w_i \vartheta_i}{D}$$

where k = 1-5, N_k = 774, and D is the number of words in the document.

In the process of calculating the five elements of the Sensory Adjectives (V, A, H, O, G), the data is normalized based on the document or chunk size so that the ratio of Richness to Sensory Adjectives becomes independent of document size. This normalised value is multiplied by its particular word weight (ϑ_i) but in this case, it is not squared by the word count frequency because there were almost seven times the quantity of these more frequently occurring prenominal adjectives to emphasize the variance in the data. The five different Sensory Adjective elements are then summed to provide an overall score, (S).

3.4.2 The CSD method

Early mentions of Critical Slowing Down (CSD) in the late 60's was attributed to the non-linear effects in Ising spin models of ferromagnets approaching the Curie point (temperature), where sharp changes in magnetic properties occur (Matsudaira, 1967). This abrupt, rapid change of state can be seen in complex systems across a variety of disciplines (ecology, climatology, finance, and medicine) as tipping-points Slater (2013). Ecosystems and biological systems are known to be inherently complex and to exhibit nonlinear dynamics, and changing system dynamics have been suggested as early warning signals (EWS) for tipping points (Scheffer, 2010; Drake & Griffen, 2010; Guttal & Jayaprakash, 2008; Barnett *et al.*, 2013; Trefois *et al.*, 2015).

While the CSD phenomenon has been observed on a number of other systems, it was originally defined in terms of the recovery time from a perturbation, and it can't easily be measured from the text. However, the definition has been broadened, and in a recent comparative study of four emotions in healthy and depressed people, van de Leemput *et al. (*2014) discovered that individuals go through a major transition in moods that are separated by tipping points. Early warning signals of such tipping-points can be detected via critical slowing down where statistical metrics such as variance, 1-lag autocorrelation, and skewness increase near the tipping point (Drake & Griffen, 2010; Slater, 2013).

Here, we ignore variance, because in ecosystems, Dakos *et al.* (2012) state that CSD variance can be systematically underestimated due to the prevalence of low frequencies close to the tipping point (it might increase or decrease, so it is difficult to test for – see Section 2.3.3), but that autocorrelation always increases toward critical transitions. Variance is also subject to bin size and initial false positives (Slater, 2013).

But in Dakos *et al.* (2008) the authors calculate the number of samples from the first one to the known CSD transition and apply a sliding window which is half that size and suggest that the window size can be varied and an appropriate one selected. This approach works when the known tipping point exists within the data. However, if this is not known, an alternate approach must be considered to eliminate the idea that we know where any event might be prior to testing. Therefore, we modify the approach taken by Dakos *et al.* (2008) and Slater (2013), and rather than have a sliding window of a fixed size, we adopt a window-fixed perspective (Foster, Bevis & Businger, 2005) where we fix the window from the first observation so that the window size increases with each iteration. This alternative technique has its own problems because of the increasing size of the data. At some point, it would eventually impact the results, but after testing with a random sample of 50 data points, we suggest having a sample size of between 19-26 (the size of our Murdoch and James data) would reasonably represents the data before any flattening would occur.

The Critical Slowing Down variables are modified from Dakos *et al.* (2008) and Slater (2013). For a given set of measurements, $Y_1, Y_2, \ldots, Y_N$ (in this case a single component of the RPAS stylometric signature of an author over a number of their works) of each at time $X_1, X_2, \ldots, X_N$, then the 1-lag autocorrelation (AR1) (see Equation 5) can be defined as follows:

Equation 5: Modified Autocorrelation at lag 1

$$AR1 = \frac{\sum_{i-1}^{N-k}(Y_i - \bar{Y})(Y_{i+1} - \bar{Y})}{\sum_{i-1}^{N}(aY_i - \bar{Y})^2}$$

where the first Y*i*, in this case, remains fixed at point 1.

For a given set of measurements, Y_1, Y_2, …, Y_N (in this case a single component of the RPAS stylometric signature of an author over a number of their works), then the Fisher-Pearson coefficient of skewness (G1) (see Equation 6) can be defined as follows:

Equation 6: Modified Coefficient of Skewness

$$G1 = \frac{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^3 / N}{s^3}$$

where $\overline{Y}$ is the mean, s is the standard deviation of the individual RPAS element used over the documents selected, N is the number of data points, and Y_i in this case remains fixed at point 1.

3.5 Seriation

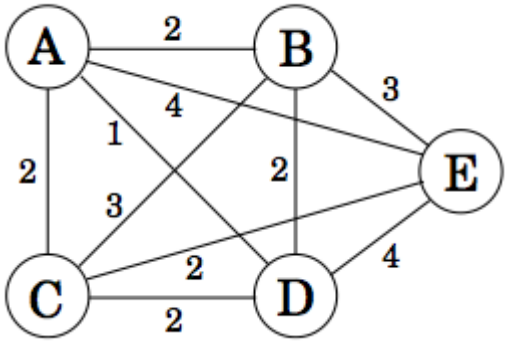
Seriation is used a number of times throughout this research thesis to support the findings of other techniques. According to Liiv (2010:71) "Seriation is an exploratory combinatorial data analysis technique to reorder objects into a sequence along a one-dimensional continuum so that it best reveals regularity and patterning among the whole series." Seriation is the process of placing a linear ordering on a set of N multi-dimensional quantities. The total number of possible orderings is N! (factorial). This grows extremely quickly with N. 5! = 120, 10! = 3.6 million and 20! = 2.4x10^{18}, or 2.4 billion billion (or quintillion). Thus, even for quite small N, we can't calculate the shortest path by calculating all possible paths. We need a heuristic or approximation. Inevitably any given approximation will work better with some data than others. Thus, for robust estimation of the shortest path, it might be necessary to try a range of different estimators and look for consistency among them.

We use the free software environment for statistical computing and graphics, R, and its seriation package (Buchta, Hornik, & Hahsler, 2008), and provide the seriation package with the matrix consisting of the nine RPAS values for each of the datasets. Using the Euclidean distance option, seriation attempts to minimise the Hamiltonian path length (the Hamiltonian path on a graph is a path which visits all the nodes just once). We

evaluated the results of the six Hamiltonian path-length calculations produced by the seriation package (TSP: *Travelling Salesperson*, Chen: *Rank two ellipse Seriation,* ARSA: *Anti-Robinson Simulated Annealing*, HC: *Hierarchical Clustering*, GW: *Hierarchical Clustering (Gruvaeus Wainer heuristic)*, and OLO: *Hierarchical Clustering (Optimal Leaf Ordering))*.

Path length calculations are complex as Figure 4 shows. If we consider each circle as a single dimensional representation of the multi-dimensional RPAS signature, each of these five documents (A-E) have a different score, and therefore the path between each of them varies. The goal of the Travelling Salesperson (TSP) approach is to travel through each document only once, and the best order of the documents is achieved through the smallest path distance. In this case, this is done for each of the RPAS elements further compounding the complexity.

Figure 4: Multi path problem



While seriation gives a one-dimensional continuum, Dendrogram branch and leaf visualization are also provided, and clusters can be separated by their Hamiltonian path distances (Earle, & Hurley, 2015). We select the technique that provides the shortest Hamiltonian path and introduce noise into the matrix to examine the strength of the connected groups by using the jitter function in R. The function adds random noise to the vector by drawing samples from the uniform distribution of the original data (Stahel & Maechler, 2011).

3.6 LIWC

The Linguistic Inquiry and Word Count (LIWC) text analysis program has been shown to be an effective tool to measure positive and negative emotion in writing (Andrei, 2014; Kahn *et al*., 2007). The emotional expression can indicate how people are experiencing the world (Tausczik & Pennebaker, 2010). Using LIWC2015, seven

linguistic emotional categories are extracted. They are emotional tone (tone), affective process (affect), positive emotion (posemo), negative emotion (negemo), anxiety (anx), anger, and sadness (sad). According to Pennebaker *et al.* (2015a), emotional tone is scored out of 100, and a high number is associated with a more positive, upbeat style, while a low number reveals greater anxiety, sadness, or hostility and a number around 50 suggests either a lack of emotionality or different levels of ambivalence The remaining six emotional categories are: Affective processes (words describing feelings) -1393 words such as affect, happy, and cried, and its sub categories: Positive emotion - 620 words such as love, nice, sweet; Negative emotion - 744 words, such as hurt, ugly, nasty; Anxiety - 116 words, such as worried, fearful; Anger - 230 words such as hate, kill, annoyed; Sadness - 136 words such as crying, grief, sad.

3.7 Mann-Whitney U Testing

We use the Mann-Whitney U-Test to compare data because it is a non-parametric independent group test to test the differences between two independent groups. This test is ideal for unequal group sizes that are small, have dissimilar variances, and a distribution that is not normal (Burns & Burns, 2012). SPSS provides three variables, U, W, and Z, but they are not overly useful. We adopt the common practice of ignoring the alternative Wilcoxon W and the Z scores given the small sample size and focus predominantly on the Asymptotic Significance (two-tailed) p-values whose rankings are generally reflected in the Mann-Whitney U scores.

The test statistic for the Mann-Whitney U Test is denoted U and is the smaller of U_1 and U_2, (see Equation 7) and is defined below:

Equation 7: Mann-Whitney U Testing

$$U_1 = n_1 n_2 + x\frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + x\frac{n_2(n_2+1)}{2} - R_2$$

where R_1 = sum of the ranks for group 1 and R_2 = sum of the ranks for group 2.

3.8 Step-wise Multiple Regression Analysis

In many multivariate situations, scientists are presented with more variables than they would like, and stepwise multiple regression discards any variables that add little or nothing to the accuracy of the correlation with the dependent variable (Beale, Kendall,

& Mann, 1967). The technique remains an effective predictive tool (Abdullah, & Rahim, 2016; Nazif *et al*., 2016; Pontone *et al*., 2016; Raghunath *et al*., 2016). The order of entry for predictors is solely a statistical decision, and here the backward entry method is adopted using SPSS (Burns & Burns, 2012). This test begins with all the predictors, in this case RPAS (including the five sensory variables V, A, H, O, and G) and removes each variable, one at a time in an iterative process and each time it accepts the most statistically significant variable that contributes to the data variability until a point where only the variables that contribute significantly are used to classify the data.

3.9 k-fold Cross-validation

In *k*-fold cross-validation, sometimes called rotation estimation, the dataset is randomly split into *k* mutually exclusive subsets of an approximately equal size known as folds (Kohavi, 1995). In this case, we use the technique with regression analysis. By setting *k* to 5, 60 samples can be split into five groups containing 12 randomly assigned samples. Conducting multiple regression analysis five times using RPAV as the independent variables, a model can be trained each time with a different fold left out so that there are five sets of unstandardized regression coefficients. The regression scores are calculated for the five folds, and a resultant accuracy scores from the average can be obtained.

3.10 Vector Space Method

In this technique, the Vector Space Method (VSM) uses both cosine and minmax methods (Koppel, & Winter, 2014; Voorhees, 1998) for similarity detection by conducting pair-wise comparisons against a single known reference data set (Koppel & Seidman, 2013) of the RPAS elements. We use the cosine and the minmax results to plot a two-dimensional array, where the closest points to the top right-hand corner are most similar, and those at the bottom left-hand corner are most dissimilar to the reference single

If we let $\vec{X} = \langle x_1, \ldots, x_n \rangle \; and \; \vec{Y} = \langle y_1, \ldots, y_n \rangle$ be the respective vector representations of documents X and Y, where each x_i represents one of the RPAS stylometric signature elements of an author's work, then (see Equation 8) cosine similarity can be defined as follows:

Equation 8: Cosine Similarity

$$sim\,(X,Y) = cosine\left(\vec{X},\vec{Y}\right) = \vec{X} * \vec{Y} / \left\|\vec{X}\right\| * \left\|\vec{Y}\right\|$$

If we let $\vec{X} = \langle x_1, \ldots, x_n \rangle$ *and* $\vec{Y} = \langle y_1, \ldots, y_n \rangle$ be the respective vector representations of documents X and Y, where each x_i represents one of the RPAS stylometric signature elements of an author's work, then (see Equation 9) minmax similarity can be defined as follows:

Equation 9: Minmax Similarity

$$sim\ (X, Y) = minmax(\vec{X}, \vec{Y}) = \frac{\sum_{i=1}^{n} \min(x_i, y_i)}{\sum_{i-1}^{n} \max(x_i, y_i)}$$

3.11 Imposters Method

This method extends the cosine and minmax approach of VSM (Seidman, 2013). We compare work that is *not* the work of the author of the referenced dataset, but an imposter. In this case, we introduce a third candidate, and the assumptions of similarity and dissimilarity are inverted. This method gives surprisingly strong results for the verification problem, even when the documents in question contain no more than 500 words (Koppel & Winter, 2014). Here, we use the cosine and the minmax results to plot a two-dimensional array, where the closest points to the top right-hand corner are most similar to the imposter and therefore dissimilar to the likely author candidate, and those at the bottom left-hand corner are more similar to the candidate.

3.12 Word Accumulation Curves

There are theoretical limits to Menhinick's Index (Equation 1) used to measure species diversity or species richness that we use above to describe Richness (R). Eventually, the value will reach the total species richness asymptote as no new species are found (Walther & Morand, 1998). In ecology, the size of the area searched impacts on the possible sample size because it is the number of species collected in a particular area and not every possible sample that exists and the measurement is species density (James & Warner, 1982). The species accumulation curve is an intuitive way to compare the richness of two samples of different sizes (Gotelli & Colwell, 2011). The species discovery curve or species accumulation curve is linked to empirical Zipf distributions (Section 3.4.1.1) can highlight differences in word frequency distribution (Bentz *et al.*, 2014).

While in Ecology, a graph where the x-axis is the number of individuals sampled, and the y-axis is the cumulative number of species recorded. Regardless of the species abundance distribution, this curve increases monotonically, with a decelerating slope

(Gotelli & Colwell, 2011). For text, the x-axis can be the document sample size, while the y-axis can be the number of unique words.

This type of curve that plots word frequency can be used to estimate the total vocabulary of a writer from a given sample (Efron & Thisted, 1976). We create two charts to examine Richness; an Accumulative Word Type Usage Curve for the largest 100 word types, and a Word Accumulative Curve. To describe the terms used, we reference Efron and Thisted (1976: 435):

> *'Note that 'type' or 'word type' will be used to indicate a distinct item in Shakespeare's vocabulary. 'Total words' will indicate a total word count including repetitions. The definition of type is any distinguishable arrangement of letters. Thus, 'girl' is a different type from 'girls' and 'throneroom' is a different type from both 'throne' and 'room'.'*

An Accumulative Word Type Usage Curve for the largest 100 word types is calculated so that we can examine the Richness of the Shakespeare and Marlowe corpus from their plotted curves using the example in Efron and Thisted (1976). Initially, we create a word type frequency list of the Shakespeare corpus and order the data from the smallest number of unique words (types) to the largest. We aggregate the data for the first 100 word groups. We do the same to the smaller Marlowe data and plot the results of both playwrights. The number of word groups (largest 100) appears on the x-axis, while the number of accumulated unique word types. We then visually compare the asymptotes of both playwrights.

A Word Accumulative Curve is calculated. Each of the works of Shakespeare is ordered from the largest work size (number of individual tokens) to the smallest. Then the number of unique words in each work (new types) introduced is calculated. This data is then aggregated, and we have a data point for each file that introduces new unique words (types). This process is also done with the works of Marlowe. We plot both playwrights. The accumulated words are written in thousands (document sample size / number of tokens) appears on the x-axis, while the accumulated unique words in thousands (number of unique words / types) appears on the y-axis.

Tweedie & Baayen (1998) have stressed lexical richness variation, and we rely on an approach using ratios (Kessler *et al.*, 1997; Singhal *et al.*, 1996) and sample below the point where the Richness variable converges. Effectively, we are examining the word density within each chunk and comparing it to the others (Gotelli & Colwell, 2011).

There were a number of techniques used which have been described above. Table 3 highlights where they were used.

3.13 Hierarchical Cluster Analysis

Cluster analysis is a type of data reduction technique linked to the concept of similarity and works by combining cases into homogeneous clusters by merging them together one at a time in a series of sequential steps (Yim & Ramdeen, 2015). We measure the distance between clusters using Ward's method. It is distinct from other methods because it uses an analysis of variance approach to evaluating the distances between clusters, and this method is very efficient and uses squared Euclidean distance to fuse cluster membership based on the smallest possible increase in the error sum of squares. SPSS provides five different types of clustering algorithms, including k-means, and Ward's method, the most popular. (Burns & Burns, 2012:553-558).

3.14 Principal Component Analysis

Principal Component Analysis (PCA) is a type of Exploratory Factor Analysis technique and aims to reduce data sets comprising a number of variables into a smaller number of datasets (called factors) that account for the underlying structure within the data (Burns & Burns, 2012: 443-450). Williams, Onsman, and Brown (2010) provide an excellent five-step guide on its conduct. However, PCA relies heavily on the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy (Kaiser, 1970) and Bartlett's Test of Sphericity (Bartlett, 1950) before the factors, expressed as eigenvalues that account for the variance in the data, can be considered.

3.15 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is used as an alternate classification technique to PCA (Balakrishnama & Ganapathiraju, 1998; Ye, Janardan, & Li, 2004). In this case, any anonymous works, or scenes are removed, and all of the individual known authors' works are numbered from 1 to N before training the model and reintroducing the unknown works (anonymous poems). Using the resultant coefficients from the Canonical Discriminant Functions, they can be aggregated, depending on how many there are above 2 to visually compare the clusters.

3.16 Receiver Operating Characterisation (ROC) Curves

Originally, receiver operating characteristics (ROC) curves were used in the detection of radar signals, but now they apply to psychology, and they are also used in medical

decision making, bioinformatics, data mining and machine learning (Robin *et al*., 2011). They have been used to depict the tradeoff between hit rates and false alarm rates of classifiers (Metz, 1978). The axis of the ROC Curve graphs is depicted by sensitivity and specificity. Fawcett (2006) highlights, that in effect sensitivity and specificity represent two kinds of accuracy: the first for actually positive cases and the second for actually negative cases. He states that one must note carefully that the terms "positive" and "negative" in these definitions concern some particular different state, which must be specified clearly in calculating and quoting sensitivity and specificity values. Metz (1978) refers to sensitivity as the proportion of correctly classified positive observations and specificity as the proportion of correctly classified negative observations. He provides the following descriptions:

Sensitivity is the true positive rate (TP Rate), also called hit rate, recall or probability of detection. Specificity is the true negative rate (TN Rate) measures the proportion of negatives that are correctly identified as negatives (See Fawcett, 2006 for a detailed explanation).

Generally, the area under the curve (AUC) measures the performance of a classifier and is frequently applied for method comparison where a higher AUC means a better classification (Metz, 1978). In this thesis, we use SPSS to construct ROC curves, and the curves are quite straight with a hard corner that reflects the Sensitivity and Specificity values.

3.17 Summary

In this chapter, we have addressed the Data and Algorithms Development phase. We have described the approach, and techniques, to be undertaken to create the author signatures so that each individual author's work can then be tested against a series of experiments. We have described how the reference data lists were created, and how the works are reduced into a Bag of Words (BOW) using the Stanford Parts of Speech Tagger. We have identified the gender (G), Richness (R), Referential Activity Power (RA), and sensory adjective (S) equations that will be used to create a signature of an author that describe self. We have restated the research hypotheses and linked each one to each of the experiments that will be undertaken.

In the next seven chapters, grouped together as Phase Two, Experiments, we will test the hypotheses and answer the research question through a series of connected studies, beginning with "Elizabethan Authorship Studies".

Part One: Elizabethan Authorship Studies

In part one of this research thesis, the focus is on identifying the authorship of a number of different works. Using 400-year-old data from the Elizabethan playwrights' and poets', William Shakespeare, Christopher Marlowe, Elizabeth Cary, Thomas Kyd, Bartholomew Griffin, Richard Barnfield, and Walter Raleigh, three studies are conducted. Each study has increasing complexity. In study one (Chapter 4) three known authors are tested. In study two (Chapter 5) one known author and 15 unknown authored play scenes are tested. In study three (Chapter 6) five known authors and 12 unknown authored poems are tested. In these three studies the first research question is addressed (Hypothesis H_1): Can a stylistic fingerprint of a person's personality – their personal signature – reveal their 'identity' from their writing style?

In this section, we have ensured that each version is free of known editorial changes that commonly occurred after Shakespeare's 1623 First Folio. However, it is not possible to know what, if any, editorial changes were made during the printing preparation and this could add some minor variation to our findings.

There are four papers that contribute to this first part (refer to Section 1.6) *Using Shakespeare's Sotto Voce to Determine True Identity from Text, Did William Shakespeare and Thomas Kyd Write Edward III?,* and *Stylometric Techniques for Multiple Author Clustering: Shakespeare's Authorship in The Passionate Pilgrim*. More detail about each one follows in the next three chapters.

Shakespeare, Marlowe, and Cary

In this chapter, the first of the three studies into the Elizabethan playwrights' and poets' identity is addressed. The study draws on the known works, and works of contested authorship, of William Shakespeare, Christopher Marlowe, and Elizabeth Cary.

Using the Stanford Parts of Speech Tagger we pre-process the data, and analysis is then conducted of the Shakespeare data. Pearson's Product Moment Correlation, r, analysis is also conducted. Using RPAS described in the methods section (Chapter 3), PtoR, AtoR, and StoR plots are constructed and analysed. Word Accumulation Curves are also constructed to support the analysis. Hierarchical Clustering Analysis (HCA), Principal Component Analysis (PCA), and Stepwise Linear Discriminant Analysis (LDA) are conducted. The findings are supported by a partial synthetic data approach.

In this initial study, the first research question is addressed (Section 1.3), and Hypothesis H_1 is tested (Section 1.4).

The findings are clear. Given Principal Component Analysis (PCA) was able to separate the known contested authored works from the known author's own works, and once the contested authored works were removed, Stepwise Linear Discriminant Analysis (LDA) was able to separate the three author's works. **Given these findings, we are able to reject the null hypothesis and say that the stylistic fingerprint of a person's personality – their personal signature – can reveal their 'identity' from their writing style**.

This chapter is taken from a peer-reviewed paper: *Using Shakespeare's Sotto Voce to Determine True Identity from Text.* It was published in Frontiers in Psychology. Vol 9. March 2018 Article 289, 1-17.

4.1 Introduction

Little is documented about Gulielmus (William) Shaksper or Shakspere, the person, outside of his christening at Stratford-on-Avon on 26 April 1564 and his marriage to Ann Hathaway in November 1582, whom he had three children with; a daughter

Susanna born in 1583 and twins, Hamnet and Judith born in 1585 (Ellis, 2000; Kreeger, 1987). However, by 1623 and seven years after his death, more than 37 plays, at least four narrative poems, and 154 sonnets had been published in London. William Shakespeare, or Shakespeare, began to be identified as the author of these works, and over the next 200 years, this solidified into a tradition (Kreeger, 1987).

Benjamin Disraeli, Lord Beaconsfield, was the first to place doubt on William Shakespeare's identity in 1837, and since then the question of the authorship of Shakespeare's publications has engaged a wide range of prominent people for more than two hundred years (Krsul & Spafford, 1997).

This ongoing controversy has engaged a lot of analysts. There are those that defend Shakespeare as author, and others whose focus is on authorship identification in general. We are in the latter group, with a focus on law enforcement identification (Kambourakis, 2014; Kaminski, 2013), and we believe this is a very fertile place to test new methods.

Although Edward de Vere, the Seventeenth Earl of Oxford has been named as a very strong candidate from a pool of fifty-six names, four major figures in English literary history, Bacon, de Vere, Stanley, and Marlowe, are the most likely alternatives to Shakespeare (Kreeger, 1987).

In 1901, Mendenhall counted the length of words and used word-length frequency distributions to separate the authored plays of William Shakespeare from Francis Bacon, and a further study found that the word-length distribution of Christopher Marlowe's plays was more aligned with Shakespeare's (Tuldava, 2004).

Elliot and Valenza (1991a) used a different identification technique and conducted modal testing based on word usage to highlight the different style of Shakespeare's poems to those of Edward de Vere and suggested that de Vere did not author the Shakespeare work.

Little is known of the creative poems of Ferdinando Stanley, also known as Lord Strange and the Fifth Earl of Derby, but he was likely associated with Shakespeare through his company of actors (May, 1972). Many believe that Shakespeare was a member of Ferdinando's acting company in the early 1590s, known then as Lord Strange's Men, before the next in line to the throne was assassinated in 1594 (Daugherty, 2011).

In 1920 doubt was raised about the authorship of the play *Titus Andronicus*, suggesting it was a pre-Shakespearian play that was retouched by Shakespeare when it was in possession of Lord Strange's men (Gray, 1920). Around the same time, Marlowe's involvement in Shakespeare's *Henry VI* was also suggested (Brooke, 1922), and today, there is still uncertainty about the influence and collaboration between Shakespeare and Marlowe (Merriam, 1998; Sawyer, 2017; Yang, Peng & Goldberger, 2017).

Other scholars have applied different techniques to the problem. Matthews and Merriam (1993) used a neural computational pattern recognition technique on Shakespeare, and fellow collaborator, playwright John Fletcher with considerable reliability, and extended their technique to the works of Shakespeare and Marlowe (Matthews and Merriam, 1994). Thirty-six Shakespeare plays, and seven Marlowe works were tested. Using ten canonical plays from Shakespeare and three of Marlowe's plays Merriam and Matthews (1994) trained their model using fifty-one one thousand word samples before subsequently classifying the remaining twenty-six entire plays of the Shakespeare First Folio and the remaining four plays from Marlowe with a success rate of 93%. They used five discriminants that comprised of a series of ratios using different combinations of the following 14 function words: *but; by; did; do; for; no; not; on; so; that; the; to; upon; with*.

In the last decade, the interest in the Elizabethan playwrights has not faded. Recent work on Marlowe and Shakespeare by Tearle *et al.* (2008) highlights that Shakespeare was a collaborator on *Titus Andronicus*, but that it was easy to separate Shakespeare from Marlowe using neural networks. Craig and Kinney (2009) suggest that there is doubt about the authorship of *Henry VI* and that Parts 1 and 2 are Marlowe's and not Shakespeare's. While Zhao and Zobel (2007) suggest that Marlowe did not write the works of Shakespeare. Our analysis suggests *Henry VI* Part I is a Shakespeare Thomas Kyd collaboration.

Stylometric analysis, the quantitative analysis of a text's linguistic features, can be traced back to Augustus de Morgan's resolution of authorship disputes using the frequency of word lengths in 1851. The first manual quantitative analysis occurred in the late 1880s by Thomas C Mendenhall (1887) who used word length distributions from the works of Bacon, Marlowe, and Shakespeare to identify the authorship of Shakespeare's plays. Stylometry has been used extensively to determine the authorship of many undocumented playwright collaborations from the Elizabethan period, including Shakespeare (Segarra *et al.*, 2015). Below we summarise some

analytical techniques, but for a more comprehensive overview of stylometry and its classification techniques see Neal *et al.* (2017) and Aljumily (2015).

Many of these stylometric text analysis techniques rely on basic statistical correlations, word counts, collocated word groups, or keyword density (Lamb *et al.*, 2013; Leech & Onwuegbuzie, 2007; Matsuo & Ishizuka, 2004). There are many different techniques in use today on Shakespeare and others. For example, n-grams (Frantzeskou *et al.,* 2007), and Latent Semantic Analysis, a method that relies on a mathematical technique called singular value decomposition to identify patterns in the relationship between terms and concepts within an unstructured mass of text (Raju *et al.*, 2016). Then there are machine learning techniques (Jockers, & Witten, 2010). However, there does not appear to be any one best technique. Juola (2008) concludes that the best choice of the feature set is strongly dependent upon the data to be analysed, and no method has yet emerged from any study as being particularly good. Rudman (1998; 2012) revisited the problem, 13 years after his earlier critic after a further 600 studies and concluded that there is still no consensus as to correct methodology or technique for authorship attribution.

There appears dissension among leading Shakespearean authorship attribution scholars about an agreed method (Rudman, 2016), but the most successful and robust methods are based on low-level information such as character n-grams or auxiliary word (function word, stop words such as articles and prepositions) frequencies (Stamatatos, 2009). The premier work in evaluating authorship in the 16th to mid-17th centuries includes MacDonald P. Jackson, Brian Vickers, and Hugh Craig (Segarra *et al.*, 2017). Jackson (2006) uses common low-frequency word phrases, repetition of phrases, collocation, and images to link word groups to other works. Vickers (2011) uses a tri-gram, or n-gram, approach, while Hirch and Craig (2014) use function word frequency. They also use methods based on the Information Theoretic measure Jensen-Shannon divergence (JSD,) and unsupervised graph partitioning clustering algorithms (Arefin *et al.*, 2015). There are other techniques used in this period of Shakespearean analysis, including simple function words (Matthews & Merriam, 1993; Merriam & Matthews, 1994) and word adjacency networks (WANs) (Segarra *et al.*, 2017), or looking at rare and unique phrases (Swaim, 2017). However, the most relevant to the RPAS technique used in this paper are the ones based on personality. The meaning-extracting method (MEM) from the field of psychology (Boyd & Pennebaker, 2015; Chung & Pennebaker, 2008) is used to extract themes from commonly used adjectives

and describe a person from their personality. Pennebaker *et al. (*2015), Litvinova *et al. (*2016) and Skillicorn *et al. (*2017) are developing personality aspects of human language to improve authorship profiling. The ability to profile user personality and infer stable differences in individual behaviour from writing can be used to predict a person's preferences and future behaviour with sufficient accuracy (Wright & Chin, 2014).

We attempt to get better clarification by going beyond statistics and blind classification and attempt to infer a person's personality; their sense of self, can be seen in subtle characteristics hidden in their writing style (Iqbal *et al.,* 2013; Argamon *et al.,* 2009; Northoff *et al.,* 2006). Voice is the manifestation of author's will, intent, and feeling; it is the animus of storytelling (Charmaz & Mitchell, 1996), an authorial voice which projects an image of the authors themselves (Lorés-Sanz, 2011). We think of this as 'sotto voce', the voice of the author that can't help but utter an involuntary truth about his identity.

Others claim to see Shakespeare's voice within his narrative. Klein (1993) says it is apparent in the guise of Hamlet's father and bound intrinsically to Shakespeare's creation. It appears in the poem, *The Phoenix and the Turtle*, as a three-part structure that foregrounds Shakespeare's voice (Cheney, 2009). It is also evident in the voice of the speaker in *The Sonnets* (Kambasković-Sawers, 2007), where "Shakespeare the man" can be reconstructed more completely here than from any of his other works (Burnham, 1990). We suggest that this voice, a person's sense of self, is reflected throughout all the works of Shakespeare, Marlowe, and Cary, and is an example of sotto voce. It can be used to determine an author's true identity.

Some of the techniques used here are not new. Richness is not, and Mendenhall used word frequency charts to separate the writings of different authors (Mendenhall, 1887). Using function words to reveal personality traits is recent but also not new (Pennebaker, 2011). Principal Component Analysis has been used extensively since the 1980's to separate the authorial styles of Shakespeare and other Elizabethan playwrights (Burrows & Craig, 2012).

However, we apply these reliable techniques to the Elizabethan playwrights to highlight the consistency of our results against other well-documented results. The creation of a stylistic fingerprint of a person from a combination of a person's internal gender, their use of sensory-based adjectives factored across the five sensory modalities, and using specific function words that have high levels of concreteness and

imagery scores which reflect self or sotto voce is new. We further highlight, how depressed a person may be from their writing. While outside the scope of this study, it is part of a broader body of work that is looking at using these techniques, particularly within the law enforcement area, where depression and the cognitive state of an individual's mental state is a valuable identifier. Using techniques that draw on biomarkers for creativity and a person's known psychological state (Zabelina *et al.*, 2015; Rosenstein *et al.*, 2015), we identify characteristics of William Shakespeare, Christopher Marlowe, and Elizabeth Cary that allow us to separate their work using a new technique RPAS.

4.2 Material and methods

4.2.1 Preparing the text

The works of William Shakespeare's *Sonnets* are drawn from the complete works of Shakespeare (Farrow, 1993), and Christopher Marlowe from Farey (2014). We also process the 1613 play, *The Tragedy of Mariam, the Fair Queen of Jewry* by English poet and dramatist, Elizabeth Cary (Mark, 2014), published after Shakespeare ceased writing, so there is an independent female writer for use in some of the testing. These versions use Modern English spelling but still contain Early Modern English words where they cannot be directly transcribed, (such as 'tis!; thou; doth, fix'd; o'er) and included for consistent word richness scores.

We divide William Shakespeare's histories, comedies, tragedies, poems and sonnets, Christopher Marlowe's plays and poems, and Elizabeth Cary's play into 57 pseudo-random textual chunks, or files (based on encountering a title heading in each work), see Table 36 in Appendix A for original and chunked data separation. This means that some chunks are partial works, such as *The Passionate Pilgrim* (chunks 23-25, and 41), *The Phoenix and the Turtle* (chunks 29-30) and *The Passionate Shepherd to His Love* (chunks 55-56). Theatrical stage direction is removed from the text (speaker titles, play actions and lists of characters for each scene) and we pre-process the files with the Stanford Parts Of Speech Tagger (Toutanova & Manning, 2000). While the tagger uses the Penn Treebank labels based on today's linguistic structure, these influences can be ignored because any variations are applied consistently across the dataset, and further they do not impact on the RPAS approach. Rather than remove all stop words – extremely common words – as is common practice, our method uses these prepositions and article word types because they carry meaning about self and a person's state of mind

that would otherwise be removed, and we only remove punctuation and symbols. The word corpus is aggregated by frequency for each chunk. We conduct an analysis of the corpus parts-of-speech tags to ensure it shows no biases and we construct a multi-dimensional vector from the results of applying **RPAS** (Section 3.4.1). While studies have successfully been conducted on one or two authors and with a single word group containing as few as 14 different words (Matthews, & Merriam, 1993, this one has three authors across a corpus of 1.031 million words and uses 507 different words in a multivariate way as described below.

4.2.2 The RPAS method

Richness is not a measure of all of the words in the English language. While the average English speaker has a passive vocabulary of about 100,000 words (Pennebaker, 2011), we are interested in Shakespeare's active vocabulary, hence limit the document size to around 30,000 words, the size of the largest Shakespeare work, rather than using smaller chunks and averaging. It should be noted that Shakespeare's Early Modern English is much closer to today's language than that of Old or Middle English and personal pronouns have maintained number, case, and gender throughout the history of English (Horobin, 2010). However, *its* only came into print in 1598, and *his* was a neuter possessive where today we would use *its*, noting that Shakespeare's First Folio, printed in 1623, kept the earlier form of *his* (Nevalainen, 2006). While we could replace *its* with *his*, there are 13 of Shakespeare's works that contain the word *its,* and we elect not to replace *his* for *its*. This approach does not affect the algorithm's effectiveness in comparing data from within the Early Modern English period. Replacing *its* with *his* would change the gender category of two poems, however, and we will mention that later.

4.2.3 Correlation Analysis

We use the SPSS (Chapman, 2017), and test the independence of the RPAS variables in the data and measure the degree of correspondence between the variables with the Pearson Product Moment Correlation or 'r' (Burns & Burns, 2012). We run three tests. In the first, we test the independence of the four high-level elements, Richness (R), personal pronouns (P), Referential Activity Power (A), and Sensory Adjectives (S). We test the sensory adjectives that make up the Sensory VAHOG elements: V – visual; A – auditory; H - haptic, O - olfactory, and G – gustatory. We also test the four linguistic variables known as particles that make up Referential Activity Power: A – articles; C –

conjunctives; P – prepositions; and PRON - pronouns. We interpret the correlation size using Burns and Burns (2012:346) descriptions.

4.2.4 Word Accumulation Curves

This type of curve that plots word frequency can be used to estimate the total vocabulary of a writer from a given sample (Efron & Thisted, 1976). We create two charts to examine Richness; an Accumulative Word Type Usage Curve for the largest 100 word types, and a Word Accumulative Curve.

An Accumulative Word Type Usage Curve for the largest 100 word types is calculated so that we can examine the Richness of the Shakespeare and Marlowe corpus from their plotted curves using the example in Efron and Thisted (1976). Initially, we create a word type frequency list of the Shakespeare corpus and order the data from the smallest number of unique words (types) to the largest. We aggregate the data for the first 100 word groups. We do the same to the smaller Marlowe data and plot the results of both playwrights. The number of word groups (largest 100) appears on the x-axis, while the number of accumulated unique word types appears on the y-axis. We then visually compare the asymptotes of both playwrights.

A different Word Accumulative Curve from the one mentioned in the previous paragraph exists, where each of the works of Shakespeare is ordered from the largest work size (number of individual word tokens) to the smallest. Then the number of unique words in each work (new word types) introduced is calculated. This data is then aggregated, and we have a data point for each file that introduces new unique words (types). This process is also done with the works of Marlowe. We plot both playwrights. The accumulated words are written in thousands (document sample size / number of tokens) appears on the x-axis, while the accumulated unique words in thousands (number of unique words / types) appears on the y-axis.

The values of lexical richness change for different measures used because of text length, and it is necessary to correct for this (Tweedie & Baayen, 1998). We do this with ratios (Kessler *et al.*, 1997; Singhal *et al.*, 1996) because we are effectively examining the word density within each chunk and comparing it to the others (Gotelli & Colwell, 2011), and any global richness coefficient can, therefore, be ignored.

4.2.5 Three Complementary Clustering Techniques

The data is clustered using three complementary techniques. The first attempts to separate the playwrights, the second separates known works from contested works – publications believed to be of different authorship – and, the third separate the three playwright's known works with the contested ones removed. SPSS is used to conduct testing.

The Hierarchical Cluster Analysis technique uses Ward's Method (Burns & Burns, 2012:557) with Squared Euclidean distance measurement, and nearest neighbour using both Squared Euclidean distance and Cosine options (see Section 3.13). The data is forced into three clusters for each playwright, Shakespeare, Marlowe, and Cary to observer where the chunks cluster.

Iterative Principal Component Analysis, PCA (Burns & Burns, 2012:443) is conducted on the known and contested works (57 chunks) to optimise the RPAS algorithm (see Section 3.4). EFA aims to reduce the variables in the data into a smaller set of factors that explain the pattern of the relationships between the variables (Burns & Burns, 2012:443). By setting the threshold to 0.30 the most non-significant RPAS variable is removed and the data retested in an iterative process until the maximum variation in the data is explained (known as the eigenvalue and it corresponds to the sum of the squared factor loadings). Once this is achieved, we use the identified components, also known as factors, for each of the significant variables that make up the components (factors) to plot the 57 chunks and observe how the known and contested works visually cluster. We test the data initially by using the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy (Kaiser, 1970) to ensure the KMO is greater than the 0.5 thresholds and deemed acceptable to continue with PCA. We also ensure that Bartlett's Test of Sphericity (Bartlett, 1950) has a significance value ($p < 0.05$) indicating there are some relationships between the variables so that PCA can extract meaningful data. We apply Kaiser's criterion rule (Kaiser, 1970) by producing a scree plot which highlights all of the eigenvalues and only retaining those factors that are above the eigenvalue of 1.

Stepwise Linear Discriminant Analysis (LDA) as an alternate classification technique to PCA is conducted (Balakrishnama, & Ganapathiraju, 1998; Ye *et al.*, 2004). We remove the contested works from the data and categorise all of the individual known authors' chunks, numbering them 1 to 3 and train the model. Using the resultant coefficients

from the three Canonical Discriminant Functions, we plot the functions and compare the clusters.

Finally, we test the effectiveness of the LDA resultant coefficients to correctly classify other works by Shakespeare. Because there is no new Shakespearian to test on, a different approach must be chosen. Rather than use k-fold cross-validation to test the accuracy of the LDA model (Rodriguez *et al.*, 2010), we draw on the full and partial synthetic data approach by Little (1993) and Rubin (1993). We elect to use the partial approach because we are not concerned with data disclosure issues (Drechsler *et al.*, 2008). Five Shakespearean works are chosen at random and divided into 2000-word chunks so that we end up with sixty-two of them. Five partially synthetic samples are constructed using 12 randomly selected chunks. These new 24,000-word synthetic works are calculated using the LDA resultant coefficients and overlayed against the uncontested works to see how close they cluster to Shakespeare, Marlowe, and Cary.

4.3 Results

Within this section, we discuss the correlation analysis results, the differences in the word accumulation curves, the hierarchical clustering, and principal component analysis. We conclude with the stepwise linear discriminant analysis predictive model that is verified using a partial synthetic approach.

4.3.1 Correlation Analysis

The independence of the variables was tested using the Pearson correlation coefficient, 'r', (see Table 38 in Appendix A) and determined for RPAS. The results were significant at the 0.01 level, with most of the relationships between the variables being deemed as weak or random (13-33%). Richness appeared to have a moderate to high correlation with Referential Activity Power, and the relationship bordered an inverse moderate to substantial level as it predicted around 69% of Referential Activity Power. In all cases, the relationship between Referential Activity Power and all other variables had an inverse relationship. Overall, the elements were independent of each other.

Pearson's correlation testing was used on the sensory adjectives that made up the Sensory element: Auditory, Gustatory, Haptic, Olfactory, and Visual. The results were significant at the 0.05-0.01 level. Of the five senses, Auditory was the weakest with either no correlation or a small random predictor relationship of 8%. Visual had the most number of correlations, but it had a weak to moderate relationship to all of the

other sensory variables (varies between 8 – 61%). Gustatory, Olfactory, and Haptic had the same correlations and did not have a significant relationship to Auditory. They also had a weak to moderate relationship to all other sensory variables (varies between 33 – 60%). Again, the elements were independent of each other.

Pearson's correlation coefficient testing was used to determine the independence of the particles that make up Referential Activity Power. The results were significant at the 0.01 level. The analysis showed that Prepositions are substantial as shown by its relationship with Articles (80.8%) and Conjunctives (73.8%) but not with pronouns (50%), and the relationship was only moderate. The correlation between Pronouns with Articles (47%) and Conjunctives (32%) highlight they were less correlated with a weak to moderate relationship. In this case, it would seem overall that the elements were less independent of each other.

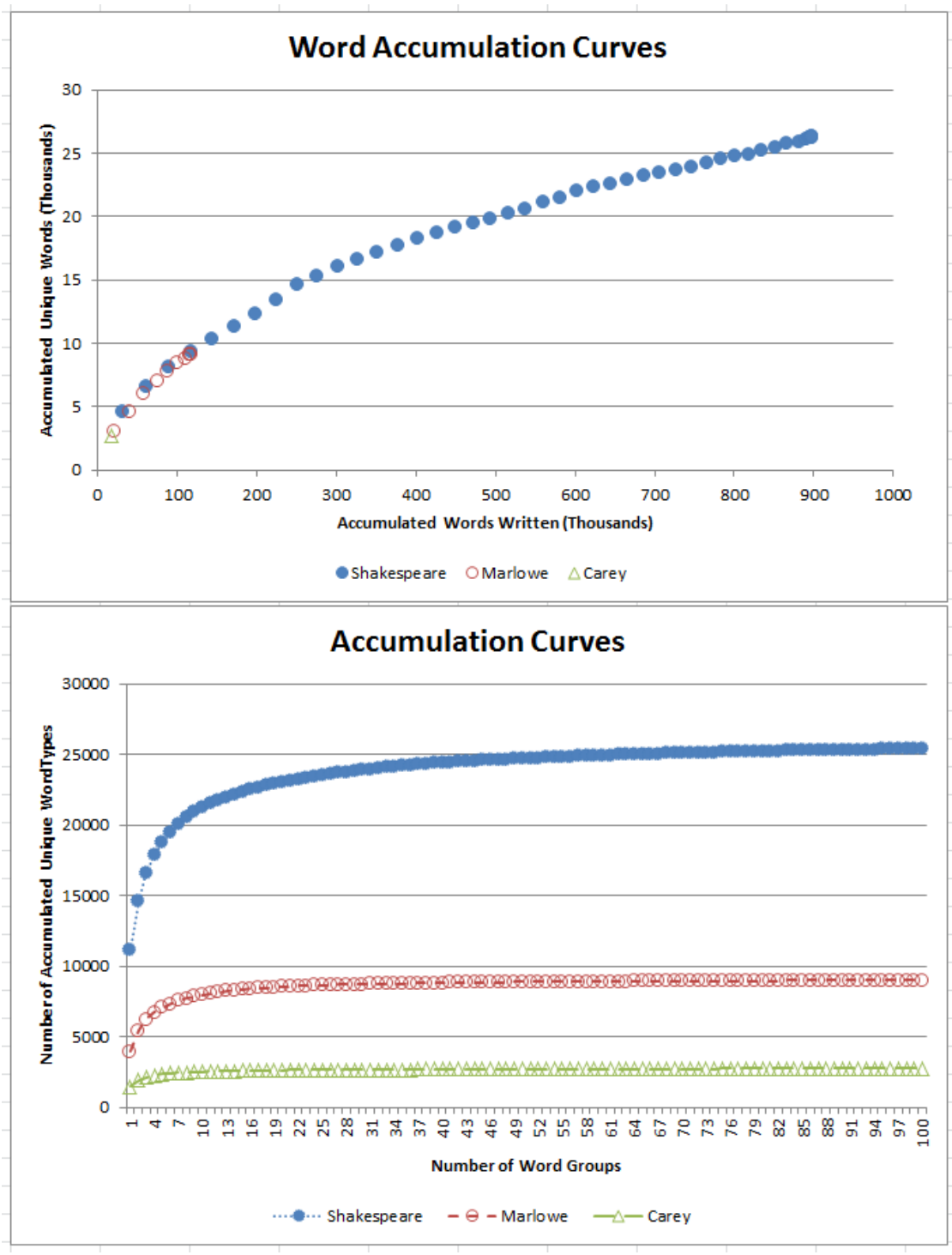
4.3.2 Word Accumulation Curves

There is a large difference in the sample sizes of Shakespeare, Marlowe, and Cary. Therefore, as an alternate test for the Richness calculations, Word Accumulation Curves were plotted for Shakespeare's 897,308-word, Marlowe's 116,446-word, and Cary's 17, 376-word corpus to examine if their use of vocabulary was similar. As can be seen (lower panel Figure 5) Shakespeare's unique word list reached an asymptote at about the 50th largest word group, which is a total of 24,726 unique words. Marlowe's unique word list reached an asymptote at about the 21st largest word group, a total of 8,565 unique words, and Cary's unique word list reached an asymptote at about the 15th largest word group with a total of 2,599 unique words. When we compared the point at where both word group curves asymptote, we could see that Marlowe used about 34.6% fewer unique words than Shakespeare, and Cary used about 89.5% fewer words than Shakespeare.

However, there is a significant difference between the number of works each produced and comparison of word accumulation plots tells a different story (upper panel Figure 5) . It highlighted that Marlowe and Shakespeare have similar word growth that might take into account the influence of vocabulary size. We cannot make a comparison with Cary with a single work. There is an age difference between Shakespeare and Marlowe which could account for these differences. People's vocabulary is known to peak late in adulthood before it declines (currently peaking around 65 years. See Hartshorne &

Germine, 2015), but this could highlight that age differences contribute to and help differentiate people from their Richness scores.

Figure 5: Word Accumulation Curves for Shakespeare, Marlowe, and Cary by groups of words the same size (word groups) and accumulated words. In the lower panel, the different number of words each playwright used (unique words) is shown and is different, but in the upper panel, the similarities between Marlowe and Shakespeare's word usage is highlighted.



Of all the works of Shakespeare over 10,000 words, the unique words contributed about 13-23% (2400-4600 words), and 45% of these words are of the small group of 450 function words that account for less than 0.1 percent of the English vocabulary but

make up more than half of the words commonly used (Pennebaker, 2011). Of all of the works of Marlowe over 10,000 words, the unique words contributed about 14-20% (2700-3200 words), and 42% are function words. In both cases, the chunks are well below a size that would approach the asymptote, and we deem that this phenomenon occurs outside of our enforced limit of a 30,000-word sample.

4.3.3 Hierarchical Clustering (HC)

To determine if there are clear differences in the writing styles of the three playwrights, the data was forced into three clusters, through Hierarchical Cluster Analysis, (using Ward's Method with Squared Euclidean distance measure, and nearest neighbour using both Squared Euclidean distance and Cosine measure). It was expected that by forcing three clusters, one for each playwright (Shakespeare, Marlowe, and Cary), they would appear in separate clusters. However, the data variations in the contested and non-contested authored works were too distant in Eucleadian space, and one of the clusters that formed had all three playwrights in them (see Table 4). Another test would need to be performed on a smaller set of the data without the contested, non-authored works, therefore as an alternative, Principal Component Analysis was conducted.

Table 4: Hierarchical Cluster Analysis Membership for 3 clusters

Cluster Membership					
Case	3 Clusters	Case	3 Clusters	Case	3 Clusters
1:C1	1	20:C7	1	39:T9	1
2:H1	1	21:C9	1	40:C14	1
3:H2	1	22:T3	1	41:P8	3
4:H3	1	23:P5	3	42:C15	1
5:H4	1	24:P4	3	43:P9	1
6:C2	1	25:P3	3	44:C16	1
7:T1	1	26:C8	1	45:C17	1
8:P1	2	27:T4	1	46:H10	1
9:C4	1	28:C10	1	47:CM1	1
10:T2	1	29:P6	3	48:CM2	1
11:P2	1	30:P7	3	49:CM3	1
12:C3	1	31:C11	1	50:CM4	1
13:C5	1	32:C12	1	51:CM5	1
14:H5	1	33:C13	1	52:CM6	1
15:H6	1	34:T5	1	53:CM7	1
16:C6	1	35:T6	1	54:CM8	2
17:H7	1	36:T7	1	55:CM9	3
18:H8	1	37:T10	1	56:CM10	3
19:H9	1	38:T8	1	57:EC1	1

4.3.4 Principal Component Analysis (PCA)

Iterative Principal Component Analysis to optimise RPAS on the basis of the maximum variance explained by eigenvalues was conducted. Initially, PCA was conducted on the four high-level variables. Only one factor was extracted and accounted for 64.3% of the variance. All the remaining three factors accounted for (35.78%) and were not significant.

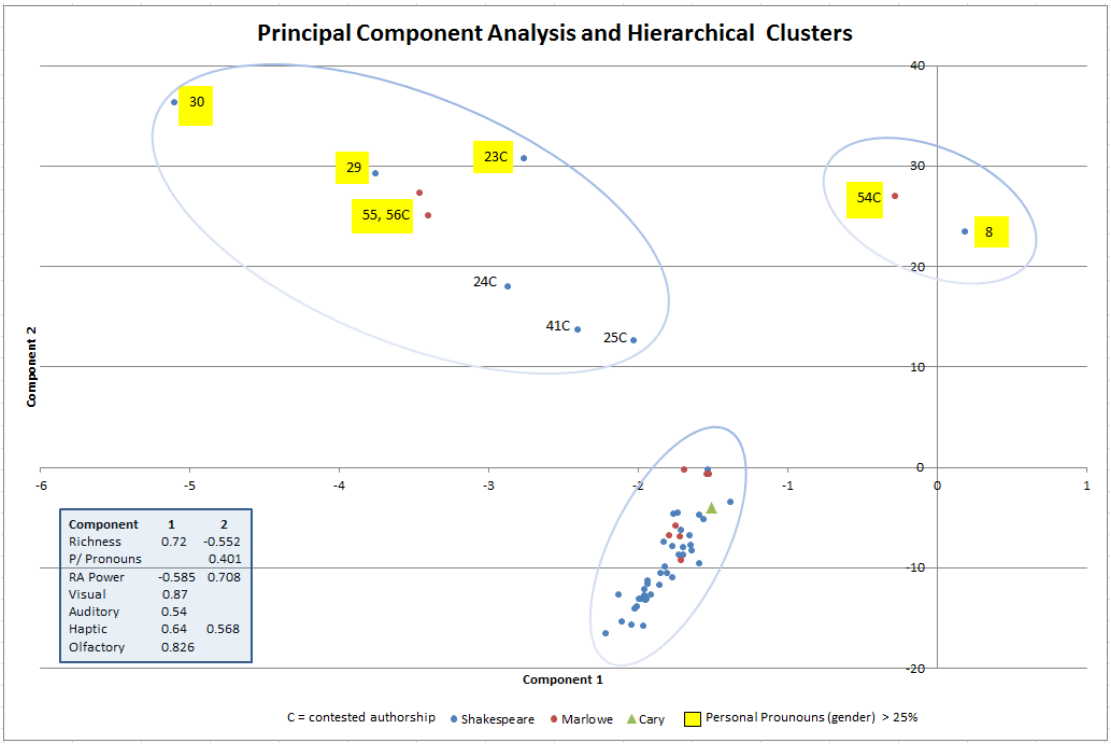
Principal Component Analysis was extended, and the Referential Activity Power element was substituted with its four variables. Articles, Conjunctives, Prepositions, and Pronouns were tested to determine if the total variance could be increased over the initial 63.4% results obtained from the single factor. However, only one factor was extracted, and it accounted for 65.6% of the variance. All the remaining six factors accounted for 34.4% and were not significant. Overall, the total variance explained by the single factor increased by 1.3% over the initial test.

Principal Component Analysis was again extended, and the Sensory element was substituted with its five variables. Now, with the Visual, Auditory, Haptic, Olfactory, and Gustatory (VAHOG) variables, many correlations were in excess of 0.30, and both the KMO and Bartlett's tests produced criteria that support the application of PCA (0.722, $p < .001$). Communalities varied from .832 (Richness) to .354 (Gender). By applying Kaiser's Rule and scree test, two factors were deemed important. Following rotation, factor one was loaded on five items that reflect four of the five sensory elements variables and RA Power accounted for 49.56% of the variance. Factor two is loaded on the Richness, personal pronouns, RA Power, and two of the Sensory adjectives (Auditory and Visual) and accounted for 22.32% of the variance. Overall, the total variance explained by the two factors was 71.88%. This is an increase of 7.6% over the initial test and 6.3% better than the second test that expanded the Referential Activity Power elements. Unweighted least squares Factor Analysis results highlighted Pearson's r correlations and indicated the inverse nature of Referential Activity Power along with the isolated Auditory variable. The Correlation Matrix, KMO and Bartlett's Test, Communalities, Total Variance Explained, and Component Matrix results are found in Table 39 - Table 45, along with the Scree Plot, Figure 36 of Appendix A.

The results of the Hierarchical Clustering and the Principal Component Analysis can be overlaid to reinforce the consistency of the results (Figure 6) and show the

separation of the contested works from the main body of works. This was identified through the two leading factors of the PCA grouped by the Hierarchical Clustering results (blue ellipses). These methods are robust enough to correlate precisely. The cluster at the bottom contains most of the chunks for all three authors. The second largest cluster on the top left contains works of uncertain or mixed authorship, such as Shakespeare's *The Passionate Pilgrim* (chunks 23-25, and 41), and Marlowe's two-authored *The Passionate Shepherd to his Love* (chunks 55-56). The exception was Shakespeare's *The Phoenix and the Turtle* (chunks 29-30). While the differences in *The Phoenix and the Turtle* have been put down to Shakespeare's genius (Bednarz, 2012) and there is still some uncertainty over authorship (Richards, 1958), it is a generally accepted Shakespearian work. The cluster on the top right showed one work each of Shakespeare and Marlowe's that is stylistically quite different from their other works (chunk 54 for example, *Hero and Leander,* was completed by George Chapman after Marlowe's death (Williams, 2005), while *Venus and Adonis* was suggested to be written during Shakespeare's hard times during the plague (Stritmatter, 2004). It is said to lack a sense of form and seen as dull (Putney, 1941). The results were reinforced by the personal pronoun analysis. Here we highlighted that most works are low in this category, and seven chunks had scores over 25% (Figure 6 yellow boxes highlight chunks 8, 23, 29-30, and 54-56). Two of these are high scores (> 80%) and appeared in the top right cluster. When comparing Richness against Referential Activity Power, four very noticeable spikes occur (chunks 24, 29-30, 41, and 55-56), and these were also the works that appear in the top left cluster. Two lesser spikes occurred in the top right cluster (8 and 54). This relationship between Richness and Referential Activity Power is unusual and is discussed further below. To further reinforce these consistent results, analysis of Richness against Sensory identified a large cluster of Shakespeare and Marlowe's works, but this time with a diffuse set of outliers. Most of these outliers were the same as those in the top clusters in Figure 6.

Figure 6: Results of the two clusters from the Principal Component Analysis overlayed with the Hierarchical Cluster Analysis results and showing the three clusters that form to separate the known works of the three playwrights from the works that are of contested authorship (or in the case of 8, 29, and 30 are stylistically different). The Personal Pronoun (gender) scores where they are > 0.25 are also shown to emphasise differences. The table highlights the contribution of the two components that the RPAS-VAHOG variables made.



4.3.5 Stepwise Linear Discriminant Analysis (LDA)

To look at the data in more detail, the contested works were removed from the data, and stepwise Linear Discriminant Analysis was conducted. LDA is better at data classification than PCA, and it is less susceptible to shape and location changes when transformed to different spaces than PCA (Ye *et al.*, 2004). The results of LDA on the eleven elements showed that three variables contributed the most to the classification of the data: Auditory, Haptic, and Richness. Two canonical discriminant functions were extracted, and both were statistically significant ($p < 0.001$, and $p = 0.008$), as was shown in the Wilks' Lambda results. The Canonical Discriminant Functions plot of each playwright also highlighted clear separation in their centroids. Using this information, we reviewed the two sensory elements, Haptic against Auditory, and Richness against Auditory to discriminate the works of each playwright. Figure 7 shows the work chunks clustered against the Auditory and Haptic sensory elements. From the group centroids, there was a clear separation of the authors. Overall, Shakespeare's chunks had a style that was higher than Marlowe in the Haptic element (0.13 vs. 0.08), and lower in Auditory (0.12 vs. 0.19) and Richness (15.5 vs. 18) with the auditory signature being a very strong separator. The LDA Eigenvalues of the first

two canonical functions, Wilk's Lambda results, Discriminant Function coefficients, group centroids are found in Table 46 - Table 49, along with the first two dimensions of a canonical discriminant analysis applied to the uncontested works of Shakespeare, Marlowe, and Cary, Figure 37 of Appendix A.

To further test the effectiveness of RPAS the new 24,000-word synthetic works were overlayed against the uncontested works. As can be seen in the Haptic and Auditory plot (Figure 7), they visually aligned closer in style to Shakespeare, and their group centroid was closer in three-dimensional Euclidean space to Shakespeare than Marlowe (a distance of 31.7 vs. 34.2).

We superimposed Richness (R), and Referential Activity Power (A)and the AtoR mapping (Figure 8 inset) highlighted several works with stylistic features likely written during difficult periods of the playwright's lives, perhaps brought about from the Bubonic Plague closing theatres, and against a backdrop of a poor economic environment and violent conditions in London during the late 1590s. The two insets highlighted a number of corresponding Richness spikes (upper diagram) with low Referential Activity Power values (chunks 8, 23, 24, 25, 41, 55, 56). These high Richness chunks were less concrete, more abstract and surreal, and they had less imagery and emotion across the sensory aspects, which highlighted a different style to the other works.

To remove any chunking bias, we resampled Shakespeare's *Venus and Adonis* and *Merchant of Venice* into 2000 word–sized chunks and plotted AtoR (Figure 8). We would have expected a lower RA Power (Bucci & Maskit, *2004*) in a depressed state, which is what we observed in the centroid differences between the two works. We see Richness as a very strong separator. However, we would also have expected to see more lexical repetition through a lower Richness score (Garrard *et al.*, 2005). It is possible that the work was an early collaboration with another author, which was why it appeared near Marlowe's collaboration with George Chapman (refer to top right cluster in Figure 6). It is also possible that the higher Richness was due to Shakespeare's large vocabulary.

Figure 7: Results of the Linear Discriminant Analysis of the uncontested works of the playwrights showing the most significant element from each canonical function (Auditory and Haptic Sensory elements). The mean of the works of each playwright is also shown. After constructing five partially synthetic Shakespeare works and overlaying them against the original data, they are closest to Shakespeare.

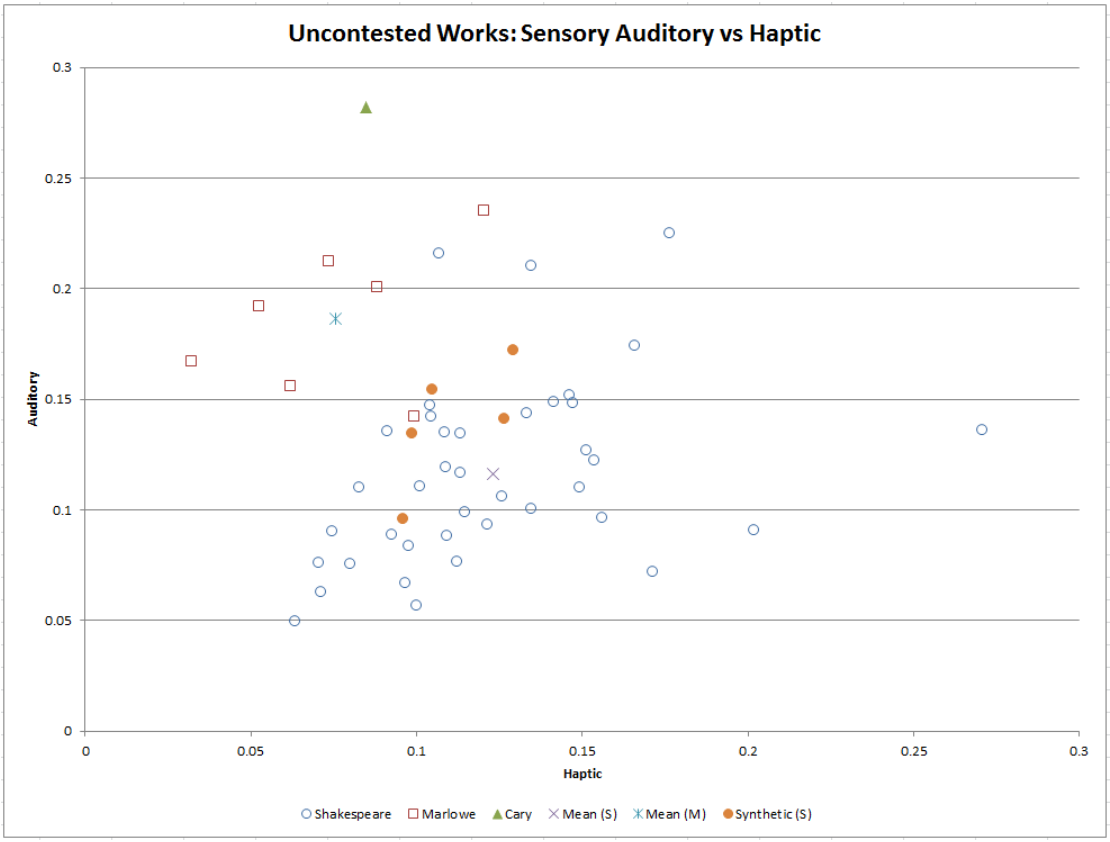
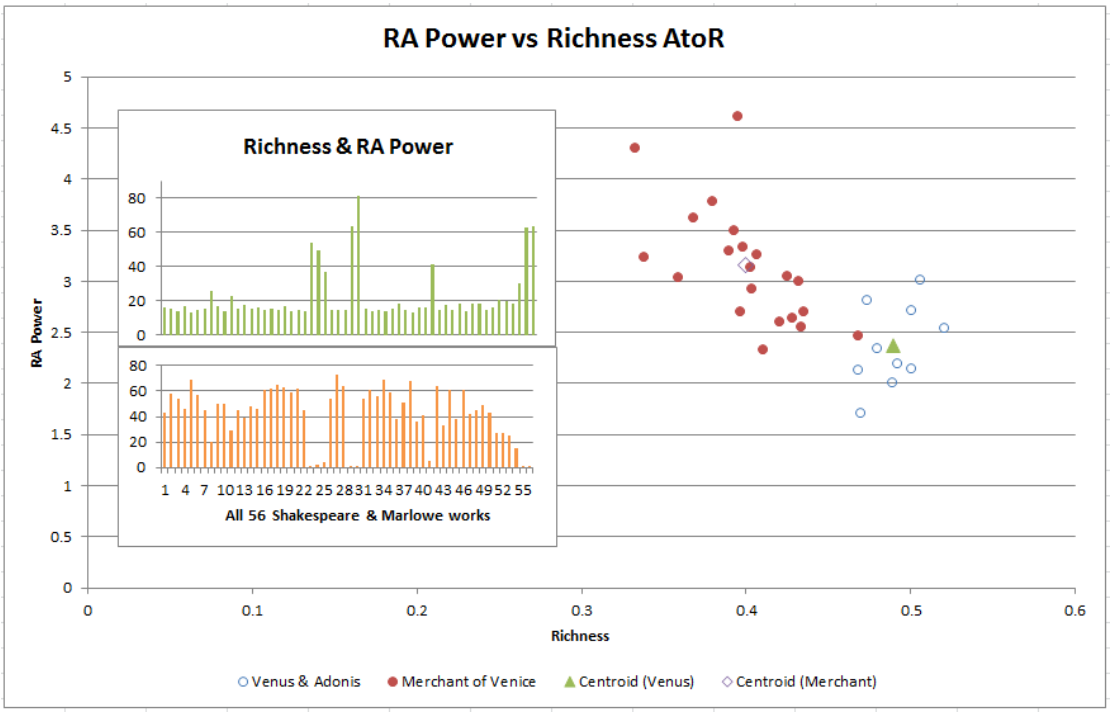


Figure 8: The Venus and Adonis play (8) which seems to be stylistically different and has an unusual Richness to Referential Activity Power relationship (see inset) is divided into 2000 word chunks as is the Merchant of Venice (16). The centroids of each play maintain the low RA Power / high Richness anomaly, highlighting the results in the inset is not an artefact of the size of the play.



4.4 Discussion

Using modern techniques on 400-year old data has some limitations. After William the Conqueror invaded England, Anglo-Norman (French) became the administrative language of Kings and nobility in England for more than 300 years. However, Anglo-Saxon (Old) English use remained in 95% of peasants and the lower class and resurged due to the 100 Year War against France, and the earlier Bubonic Plague in the mid-fourteenth century. Shakespeare's Early Modern England emerged, borrowing over 10,000 Norman words, removing noun genders, and simplifying adjective inflections. The Great Vowel Shift commenced (Mastin, 2011) and pronunciation changed during 1350 to 1700. It marked the point at which language became more standardised and akin to today.

To further put the results into perspective, Early Modern English began around the sixteenth century when vocabulary expanded at its greatest rate, and it is much closer to today's language than that of Old or Middle English (Horobin, 2010). By this time pronouns, *they*, *their*, *them* had become firmly established in the standard language, such as most personal pronouns that have maintained number, case, and gender throughout the history of English. The word *its* only came into print in 1598, and *his* was a neuter possessive where today we would use *its* (Nevalainen, 2006). While we elected not to replace *its* with *his* words because while *its* does not appear in any copy of Shakespeare's works published during his lifetime, some instances do appear in his posthumous published plays. Replacing *its* with *his* would change the gender category of two poems, *A Lover's Complaint* (personal pronouns score moved from .03 to .96) and *The Rape of Lucrece* (personal pronouns score moved from .003 to 1). While *A Lover's Complaint* has been attributed to the poet John Davies of Hereford by Brian Vickers (2014), Wilson (1988) says that *The Rape of Lucrece* occupies an uncertain position in Shakespeare's canon, as an early, apprentice, experimental piece. Our analysis before using the word *his* instead of *its* suggests that outside of the higher gender score from personal pronoun use, *The Rape of Lucrece* is a Shakespeare written poem, while *A Lover's Complaint* was a contested work not written by Shakespeare. Distinct sets of indefinite and definite articles and demonstratives also existed by this time and support our algorithm's success to define the self from RA Power also, any many of the 117 function words taken from the MRC Psycholinguistic database were used during this period. While the meaning of some words has changed over time, many of the

sensory adjectives from the list were not identified, but there were enough early and simpler Early Modern English words identified to be of value.

Empirical Zipf distributions and word accumulation curves have been used to highlight differences in word frequency distribution between Old English and Modern English of about 23%, whereas the differences between Early Modern English and Modern English is around 10% with the two modern language distributions being similar in terms of case, marking, and other inflectional paradigms like subjunctive ones, which have been replaced today by modal verbs (Bentz *et al.*, 2014). Language does change over time, as does the meaning of some words, but by applying our approach across all of the Elizabethan works only and not drawing on any modern English works, any bias is consistent and does not change the clustering results.

Estimating Shakespeare' word use for authorship identification purposes might be effective (see the Taylor poem in Thisted & Efron, 1987). It is known that Shakespeare had an active vocabulary of over 21,000 different words, and while today's educated person's vocabulary is less than half that, Shakespeare has been credited with introducing more than two thousand words into today's everyday use (Bragg, 2003). Shakespeare's strength was his support from the King, to write and perform his plays in the emerging trade centre, London for all to hear, the impact akin to today's newspapers and the internet. Brown and Gilman (1989) suggest that Shakespeare's dramatic text provide the best information on the colloquial speech of the period. He represented the conduct within court and society during a rich period of cultural reform and loaned from a library of lost voices (Bristol, 1996). Shakespeare's works are overrepresented in the first edition of the Oxford English Dictionary, contributing almost 33,000 quotations (Hoffmann, 2004), and he would have leaned on existing words in use during this important period of language reform. Notwithstanding this, it was estimated that Shakespeare knew an additional 35,000 words he did not use (Efron & Thisted, 1976). Word accumulation curves (Figure 1) highlighted that during his life Shakespeare used around 21% more unique words than Marlowe. However, there was a significant difference between the number of works each produced and comparison of word accumulation plots highlight they have similar word growth that might take into account the influence of vocabulary size varying with age differences (Hartshorne & Germine, 2015). Regression Analysis showed similar Richness characteristics for Shakespeare and Marlowe, and results of two-sample T-Tests (p-value 0.980) also suggested no significant difference between Shakespeare and Marlowe when Johnson

Arcsine Transformations are applied to normalise the positively skewed data. Therefore, we suggest Richness (R) is a valuable stylistic contributor for authorship identification.

The correlation analysis of the four high-level RPAS variables highlighted that the RPAS variables are best used in this configuration, or as RPAS (VAHOG) without the five independent sensory elements aggregated into one Sensory Adjective (S) variable. This was also highlighted in the results of the Linear Discriminant Analysis.

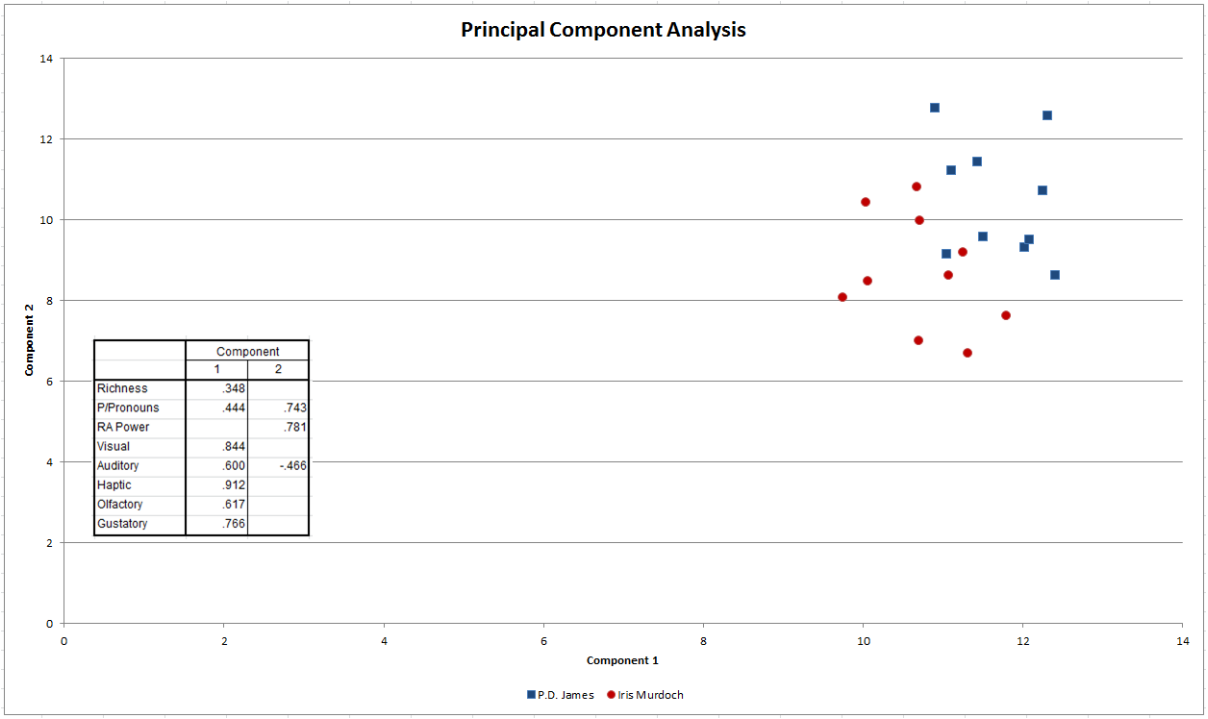
There were also some periods where there seemed to be 'depression-like' episodes where RA Power dips predominantly (as shown by AtoR in Figure 8). These results are also reflected in the sensory-based adjectives and might be useful in determining changes in the cognitive states of people. Further analysis would need to be conducted using recent data of depressed subjects. However, it has the potential to identify characteristics of self within cyberspace for law enforcement purposes.

4.5 Testing the PCA and LDA Concepts on Contemporary Data

There is no doubt that the Shakespearian dataset is not sufficient to rely on solely. Therefore, we draw on contemporary data from research conducted in later chapters (Chapter 8 and 9) to emphasise the possible connection to low RA Power scores. Using the techniques from the first study of the Elizabethan playwrights and poets, Principal Component Analysis (PCA) and Stepwise Linear Discriminant Analysis (LDA) is conducted on this data to demonstrate that the techniques can separate the writing of contemporary authors and not only 400-year-old text. Ten samples each from novelists Iris Murdoch and P.D. James were selected, ensuring that these were not at a point where Iris Murdoch's markers for Alzheimer's disease (AD) has manifested and skew the findings.

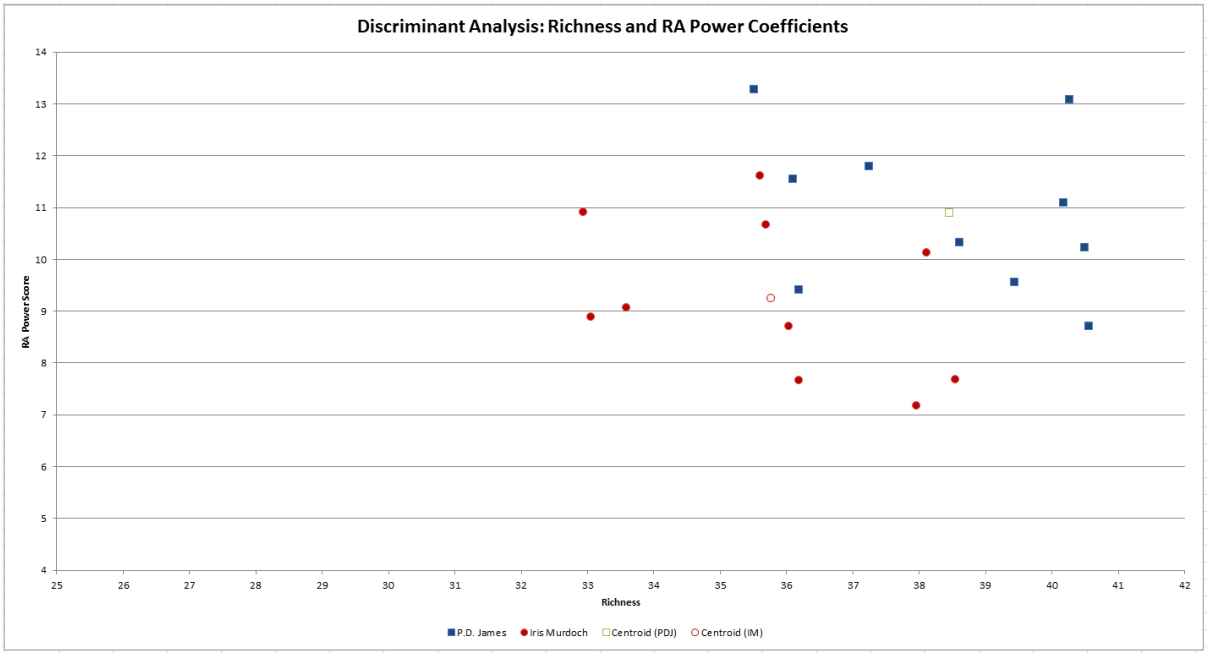
As can be seen from the Principal Component Analysis in Figure 9, two components were sufficient to separate the stylometric signatures of both authors. In conducting Linear Discriminant Analysis, Richness and Referential Activity Power, the two most significant elements from each canonical function are used. The means of the ten novel samples are highlighted in Figure 10 to suggest that Iris Murdoch's Referential Activity Power and Richness are lower than P.D. James. Murdoch's writing throughout her life did indicate a trend of falling Richness values in her novels earlier than a period 12 years before her formal diagnosis of AD, but a number of her works were quite high.

Figure 9: Results of ten samples each from the authors Iris Murdoch and P.D. James, showing the different groupings from the Principal Component Analysis. These were early samples for both authors to ensure that the impacts of Murdoch's Alzheimer's disease did not increase the variance of her sample. The table highlights the contribution of the two components that the RPAS-VAHOG variables made.



	Component	
	1	2
Richness	.348	
P/Pronouns	.444	.743
RA Power		.781
Visual	.844	
Auditory	.600	-.466
Haptic	.912	
Olfactory	.617	
Gustatory	.766	

Figure 10: Results of the Linear Discriminant Analysis of the ten samples each from the authors Iris Murdoch and P.D. James showing the most significant elements from each canonical function (Richness and Referential Activity Power). A comparison of both means of the ten novel samples is also shown and highlights that Murdoch's Referential Activity Power is lower (more depressed) with a lower Richness score, which would indicate there are still markers for depression / anxiety in the samples.



There were also three documented periods in her life where her Referential Activity Power scores fell below 10 and matched times of her known depression. This could

suggest that Murdoch's writing has markers for depression, as indicated by the lower Referential Activity Power scores, but we leave that idea for further discussion in Chapter 8.

4.6 Conclusions

We find RPAS is a different approach to the identification of self. It includes words that are strong in concreteness and imageability that reflect known psychological states in an individual's personality. The use of "sotto voce", the authorial voice which projects the true identity of the authors has enabled us to separate Shakespeare's works. Using RPAS and the PCA and LDA techniques on contemporary authors, we were also able to separate the authorship of Iris Murdoch and P.D. James. The broader implications of this research may provide signalling of depressive episodes that could have major social implications, such as averting suicide.

4.7 Summary

In this chapter, the use of RPAS equations was found to be effective at identifying the authorship of the Elizabethan playwrights. The most significant findings for this chapter were that PCA was able to separate the contested authored works from the known works of the authors using RPAS and that the LDA technique was able to differentiate the known authored works using the Sensory Haptic and Auditory elements of RPAS.

Edward III

In this chapter, the second of the three studies into the Elizabethan playwrights' and poets' identity is addressed. The study draws on the known works of William Shakespeare and Thomas Kyd. The study aims to identify the authorship of the 19 unknown authored play scenes from the unknown authored play, *The Reign of King Edward III* using RPAS.

Using RPAS (Chapter 3 methods), PtoR, AtoR, and StoR plots are constructed and analysed. The Vector Space Method (VSM) is conducted to cluster the unknown scenes, along with its variant, The Imposter's Method. The findings are supported using Seriation with noise.

In this second study, the first research question is again addressed (Section 1.3) and Hypothesis H_1 (Section 1.4) is retested using more complex data.

The Vector Space Method (VSM) and alternate Imposters Method were supported by the Seriation with noise technique and the PtoR and StoR clustering. They consistently separated works by Shakespeare and Kyd. **Given these findings, we are able to reject the null hypothesis and say that the stylistic fingerprint of a person's personality – their personal signature – can reveal their 'identity' from their writing style.**

This chapter is taken from a peer-reviewed paper: *Did William Shakespeare and Thomas Kyd Write Edward III?* It was accepted for publication by the *International Journal on Natural Language Computing*. Vol. 6, No. 6. December 2017.

5.1 Introduction

The Reign of King Edward III (*Edward III*) play was first published in 1596 and is of uncertain authorship (Slater, 1988). However, it is a new addition to the Shakespeare canon, and even while there is a suggestion that Shakespeare is not the sole author, he is considered to be a significant one (Shakespeare, & Melchiori, 1998). It wasn't included in the selection of traditional works used in Chapter 4. However, the idea that the work might be Shakespeare's was first suggested by Edward Capell in 1760 (Champion, 1988). Many others have offered their own candidate from the list of

popular playwrights of the time since then, including that the 19 scenes within it are all Shakespeare's, but more recently, Brian Vickers, using plagiarism detection software, has suggested that Thomas Kyd is the major author with Shakespeare having a lesser role (Vickers, 2014). We test the claim that the anonymous play, *Edward III,* was co-written by William Shakespeare and Thomas Kyd.

5.1.1 An approach

While many text analysis techniques rely on basic statistical correlations, word counts, collocated word groups, or keyword density (Lamb, Paul, & Dredze, 2013; Leech & Onwuegbuzie, 2007; Matsuo & Ishizuka, 2004), Vicker's approach using trigrams, instances where three consecutive words in a sentence closely match known authored works, is a more recent technique.

We use **RPAS** to create stylistic signatures of Shakespeare, Kyd, and Marlowe and compare them to the 19 scenes within the *Edward III* play to suggest the authorship before comparing these results to the recent study by Vickers. We also label the four commonly understood scenes attributed to William Shakespeare (scenes 1.2, 2.1, 2.2. and 4.4 which we refer to as chunks 2, 3, 4 and 13). We then analyse and cluster the results to identify the likely authorship of the 19 unknown scenes within *Edward III.*

5.2 Methodology

We draw on the June 1999 Project Gutenberg Etext of *The Reign of King Edward the Third*, attributed in part to William Shakespeare, and the February 2011 Project Gutenberg EBook of *The Spanish Tragedy*, by Thomas Kyd. While scholars have argued that Shakespeare's writing can be seen in the additional passages of Thomas Kyd's fourth quarto of *The Spanish Tragedy (*Bruster, 2013), we have used an earlier version, the second quarto printed in 1592 to avoid any influence of Shakespeare in the results. Consideration was given to using additional works from Kyd. *Cornelia* was discarded because it is a translation of a known earlier work of another author. The anonymous play *Soliman and Perseda* is now being attributed to Kyd because it is presented as a summarized plot in *The Spanish Tragedy*, however, features within the play have also been attributed to Shakespeare, Marlowe, and Kyd (Merriam, 1995) and it too was discarded in favour of *The Spanish Tragedy.*

We remove all stage direction and pre-process both files with the Stanford Parts Of Speech Tagger (Toutanova & Manning, 2000) to remove all stop words and then

aggregate the works by word frequency. *The Reign of King Edward the Third* is further broken down into segments marked by the 19 scenes (average scene size of 1035 words, 88 – 3720).

We use William Shakespeare's *Venus and Adonis*, and Christopher Marlowe's *Hero and Leander,* which are drawn from a pre-processed corpus (see Chapter 4) and originally sourced from the complete works of Shakespeare (Farrow, 1993) and Marlowe (Farey, 2014). We norrmalised the RPAS Personal Pronouns (P) across the 21,300 words from Kyd's *The Spanish Tragedy*, the 19,600 words in the anonymous *Edward III* play, and divide them by the 897,000-word corpus of Shakespeare and Marlowe so that all of the Personal Pronoun results are consistent across all tests.

We create a nine-dimensional array from the data using RPAS and apply three complementary techniques to reduce any single bias and determine the possible authorship of the 19 scenes. As a final measure, we then use seriation to visualise the nine-dimensional array as a one-dimensional continuum and get some distance metrics between the clusters, before adding noise to test the strength of the co-located cluster edges.

5.2.1 Three Complementary Techniques

Apart from examining the individual RPAS values from the 19 *Edward III* scenes (which includes the Sensory sub-elements VAHOG Visual, Auditory, Haptic, Olfactory, and Gustatory measures), we plot Personal Pronouns (P) against Richness (R) (PtoR), Referential Activity Power (A) to Richness (R) (AtoR), and Sensory Adjectives (S) to Richness (R) (StoR) and examine the clusters that form.

We then use the Vector Space Method (VSM) technique (Koppel, & Winter, 2014; Voorhees, 1998), see Section 3.10 for more details on the technique. We conduct pair-wise comparisons of each of the 19 *Edward III* scenes against Thomas Kyd's work, *The Spanish Tragedy* and William Shakespeare's poem *Venus and Adonis* (each a 36 pair-wise comparison) using both cosine and minmax similarity detection (Koppel & Seidman, 2013). and plot them to examine the clusters.

Extending the cosine and minmax approach of VSM, we then use the imposter's method (Seidman, 2013), where we compare work that is *not* the work of either of the two authors and in this case, is used to cluster commonly authored scenes (see Section 3.11). This method gives surprisingly strong results for the verification problem, even

when the documents in question contain no more than 500 words (Koppel & Winter, 2014). We select Christopher Marlowe's poem, *Hero and Leander*, completed by George Chapman after Marlowe's death which is very stylistically different to both Kyd and Shakespeare's work.

5.2.2 Seriation

We provide the R seriation package with the 9x19 matrix consisting of the nine RPAS values for each of the 19 scenes of *Edward III*. A detailed description of the seriation method can be found in Section 3.5.

5.3 Results

Discriminating by sensory-based VAHOG measures having zero scores, chunks 8, 10, 11, 12, 14, 15, 16, 17 stand out, and we include chunks 6 and 18 to the list when we consider Richness and Personal Pronouns scores (R > 50 or P > 0).

We conduct visual clustering and plot Personal Pronouns (P) against Richness (R) (PtoR) (see , Referential Activity Power (A) against Richness (R) (AtoR), and Sensory Adjectives (S) against Richness (R) (StoR) and examine the results (see Figure 11 (PtoR) and Figure 12 (StoR). AtoR is omitted here because it mimics PtoR, (see Figure 38 in Appendix A). PtoR discriminates chunks 6, 8, 10, 11, 15, 16, 17, and 18 by Richness and Personal Pronouns. Of these, chunks 6, 8, 15, 16, 17 have a richer and less feminine gendered style (R > 50 and P > 8) while chunks 10, 11, and 18 appear ambiguous. AtoR reinforces the Richness aspects but does not contribute further (see Figure 38 in Appendix A). StoR supports the PtoR results and discriminates chunks 3, 6, 8, 10, 12, 15, 16, 17, and 18 by Richness and a wide sensory range. Of these, chunks 6, 8, 15, 16, and 17 have a richer and much wider sensory range (where R > 50), while chunks 10, 12, and 18 appear ambiguous.

Figure 11: In this *Edward III* gendered Personal pronouns (P) versus Richness (R) diagram we see chunks 6, 8, 10, 11, 15, 16, 17, and 18 with greater than 50% richness or greater than 0 gendered personal pronouns. Of these, we see the 'Shakespeare 5' in the smaller shaded ellipse having higher female gender scores and Richness, while 18, 10, and 11 might be ambiguous. Of note, the four Shakespearian clusters marked with a red circle are those commonly attributed to Shakespeare. Further, the ellipses are our visual clustering assignment.

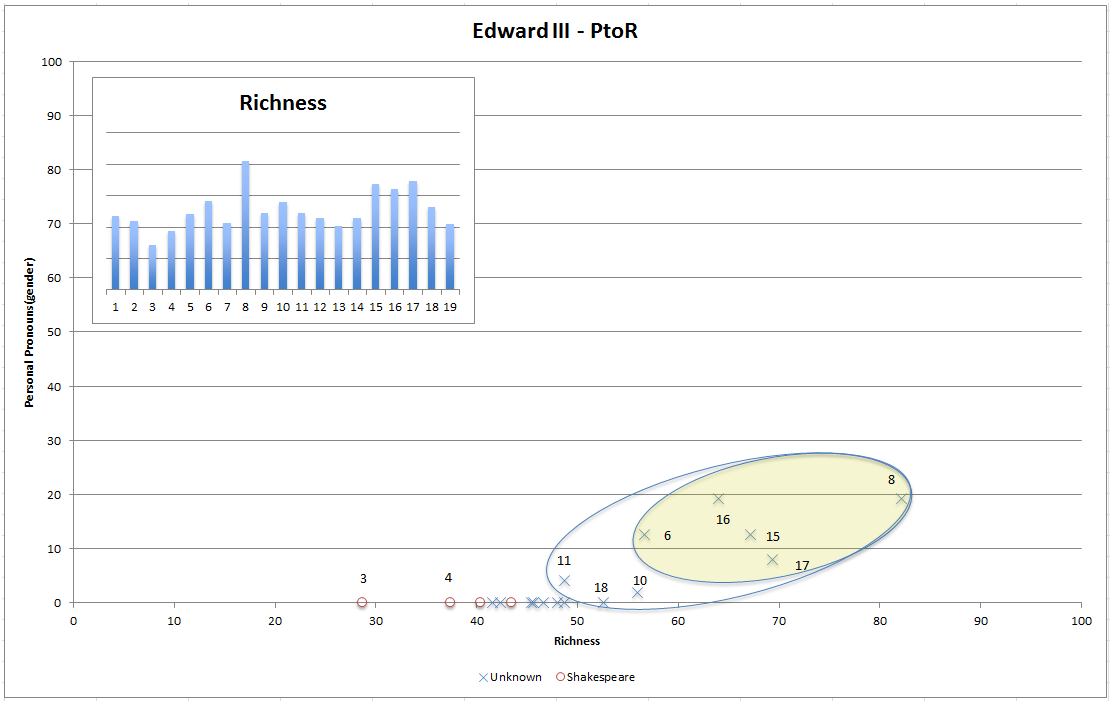
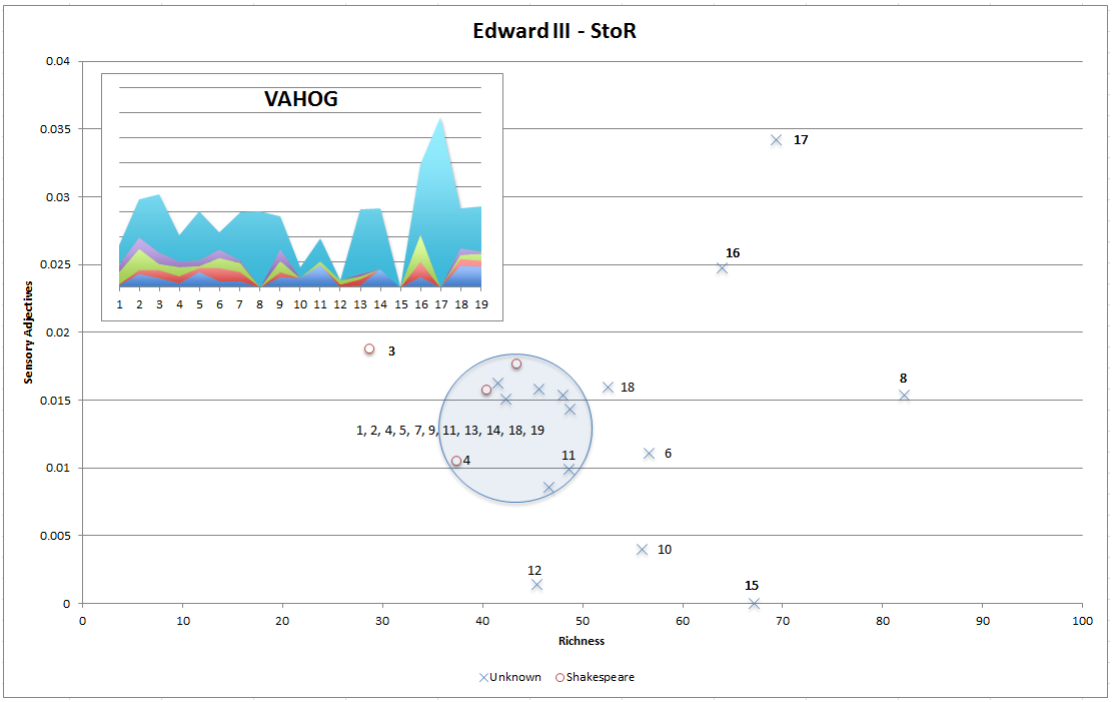


Figure 12: In this *Edward III* Sensory Adjectives (S) versus Richness (R) diagram with the VAHOG sensory allocations, we see chunks 3, 6, 8, 10, 12, 15, 16, 17, and 18 stand out because Richness is greater than 50% or they have a wide sensory range when compared to the shaded circle Kyd chunks. Of these, we see the 'Shakespeare 5' (6, 8, 15, 16, and 17) having a much wider sensory range with higher Richness, while 18, and 10 might be ambiguous. Of note, the four Shakespearian clusters marked with a red circle are those commonly attributed to Shakespeare. Further, the shaded circle is our visual clustering assignment.

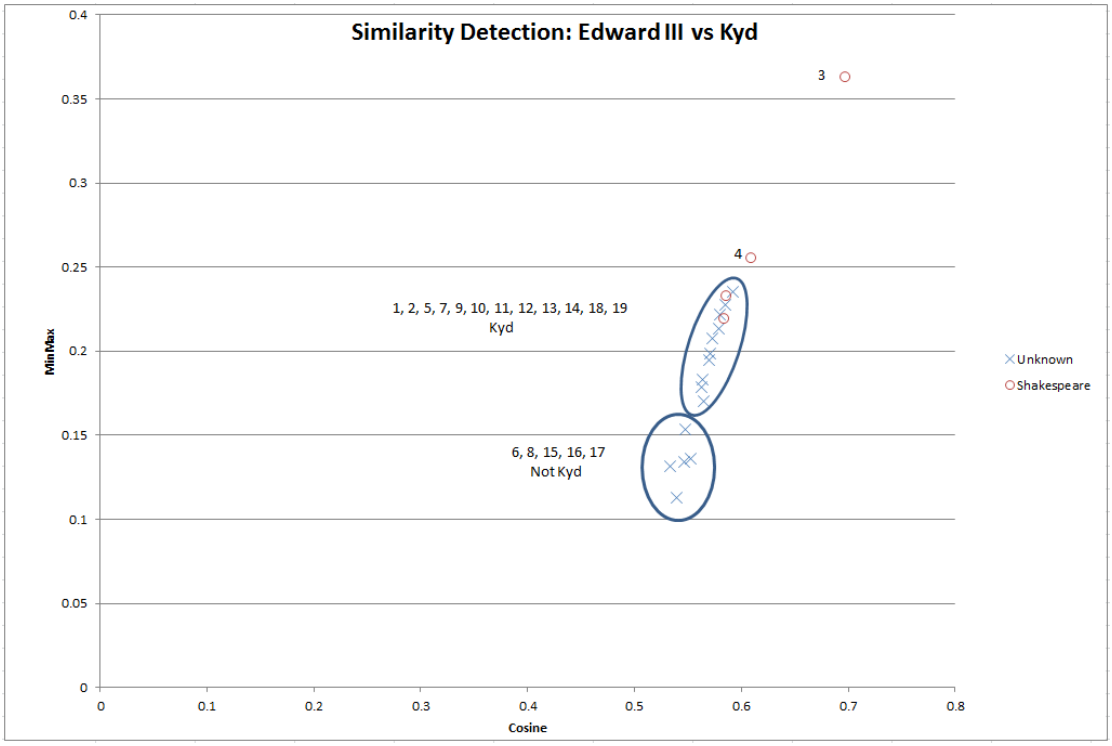


The cluster assignment is discretionary, however, through amalgamation these techniques identify two groups (Group 1: 1, 2, 3, 4, 5, 7, 9, 13, 14, 19 and Group 2: 6, 8, 10, 11, 12, 15, 16, 17, 18) with some variation in chunks 10, 11, 12, and 18.

5.3.1 Vector Space Method (VSM)

We create a stylistic signature of Thomas Kyd's play, *The Spanish Tragedy* using the **RPAS** method, and compare it to the 19 *Edward III* scenes using both cosine and minmax similarity detection (36 pair-wise comparisons). We plot these as an XY Cartesian product in Figure 13 and examine the clusters. We expect Kyd's authored chunks to appear in the upper-right-hand corner (a larger value indicates the scene is more similar to Kyd), and ones furthest away (bottom-left-hand corner) to be Shakespeare's works. The similarity plot (Figure 13) highlights two clusters, and we assign Kyd's authorship to one (chunks 1, 2, 5, 7, 9, 10, 11, 12, 13, 14, 18, 19) and Shakespeare to the other (chunks 6, 8, 15, 16, 17). Chunks 3 and 4 sit outside but also indicate Kyd.

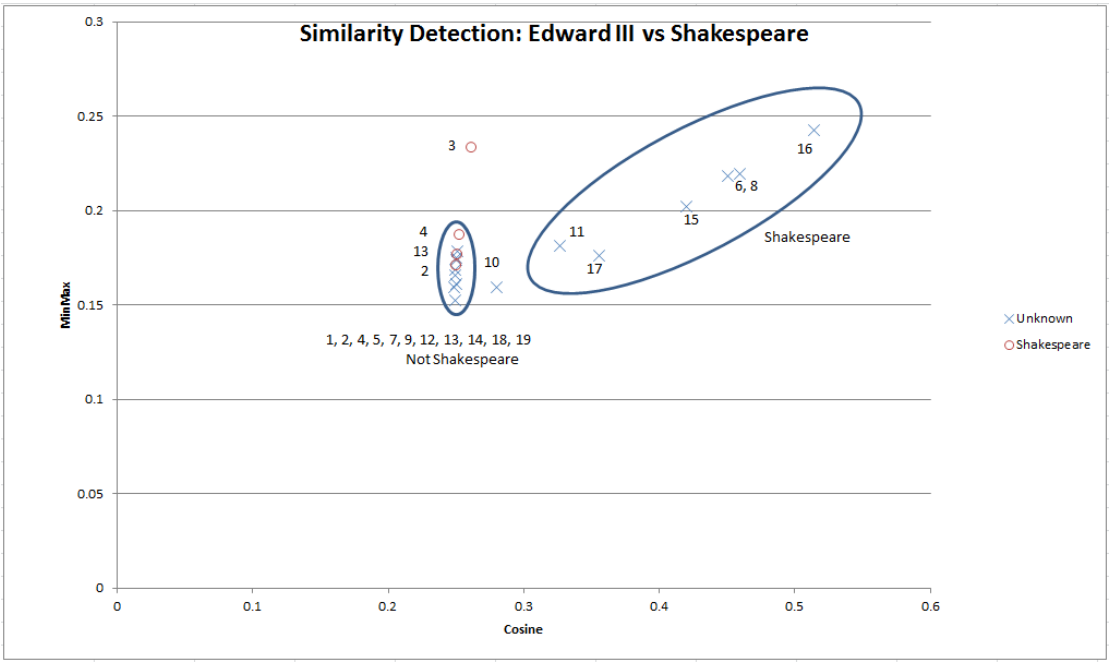
Figure 13: Using the 9-dimensional RPAS vector we compare Thomas Kyd's *Spanish Tragedy* to *Edward III* scene chunks using minmax and cosine similarity detection. We see the extreme values of chunks 3 and 4(commonly attributed to Shakespeare), and these are clearly Kyd on this metric. The 'Shakespeare 5' (chunks 6, 8, 15, 16, 17) appears in the lower cluster. Of note, the four Shakespearian clusters marked with a red circle are those commonly attributed to Shakespeare. Further, the ellipses are our visual clustering assignment.



Next, we compare Shakespeare's *Venus and Adonis* to Thomas Kyd's *The Spanish Tragedy* using both cosine and minmax similarity detection. The similarity between

Shakespeare and Kyd's work is small (cosine 21.96%, minmax 18.29% or ~79.9% dissimilar) so Shakespeare's authored chunks will sit closer to the upper-right-hand corner. The similarity plot (Figure 14) highlights a Shakespeare cluster (chunks 6, 8, 11, 15, 16, 17), and a Kyd cluster (chunks 1, 2, 4, 5, 7, 9, 12, 13, 14, 18, 19). Chunks 3 and 10 sit outside and are more like Kyd on the cosine measure, but ambiguous on the minmax measure.

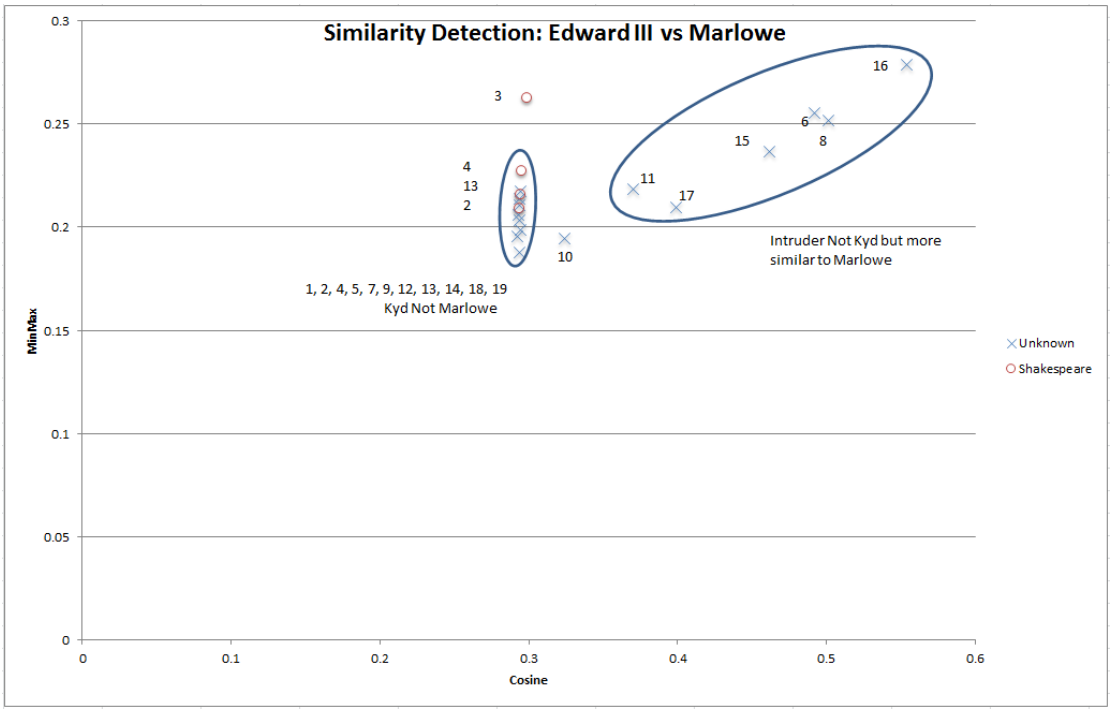
Figure 14: Using the 9-dimensional RPAS vector we compare William Shakespeare's *Venus and Adonis* to *Edward III* scene chunks using minmax and cosine similarity detection. We see the 'Shakespeare 5' (chunks 6, 8, 15, 16, 17) appear in the top cluster closest to Shakespeare's work, but with the inclusion of cluster 11. The cluster in the lower left corner clearly highlights Kyd's work as different from Shakespeare. Of note, the four Shakespearian clusters marked with a red circle are those commonly attributed to Shakespeare, and none falls close to Shakespeare. Further, the ellipses are our visual clustering assignment.



5.3.2 Imposters method

We compare Christopher Marlowe's *Hero and Leander* to Thomas Kyd's *The Spanish Tragedy* using both cosine and minmax. Marlowe's work is considered to an imposter because he is neither Kyd nor Shakespeare. The similarity between Marlowe and Kyd's work is small (cosine 22.096%, minmax 17.48% or ~80.2% dissimilar) so Kyd's authored chunks will sit furthest from the upper-right-hand corner. The similarity plot (Figure 15) highlights a Shakespeare cluster (chunks 6, 8, 11, 15, 16, 17), and a Kyd cluster (chunks 1, 2, 4, 5, 7, 9, 12, 13, 14, 18, 19). Chunks 3 and 10 sit outside and are more like Kyd on the cosine measure, but ambiguous on the minmax measure.

Figure 15: Using the 9-dimensional RPAS vector we use the Imposter Method and compare Christopher Marlowe's *Hero and Leander* to *Edward III* scene chunks using minmax and cosine similarity detection. Marlowe's work is dissimilar to Kyd's, and therefore, the work furthest away from Marlowe's is Kyd's. Logically, if there are only two authors in Edward III, then the work closest to Marlowe must be Shakespeare. We see the 'Shakespeare 5' (chunks 6, 8, 15, 16, 17) appear in the top cluster closest to Marlowe's work, but with the inclusion of cluster 11. Of note, the four Shakespearian clusters marked with a red circle are those commonly attributed to Shakespeare, and they all fall close to Kyd. Further, the ellipses are our visual clustering assignment.

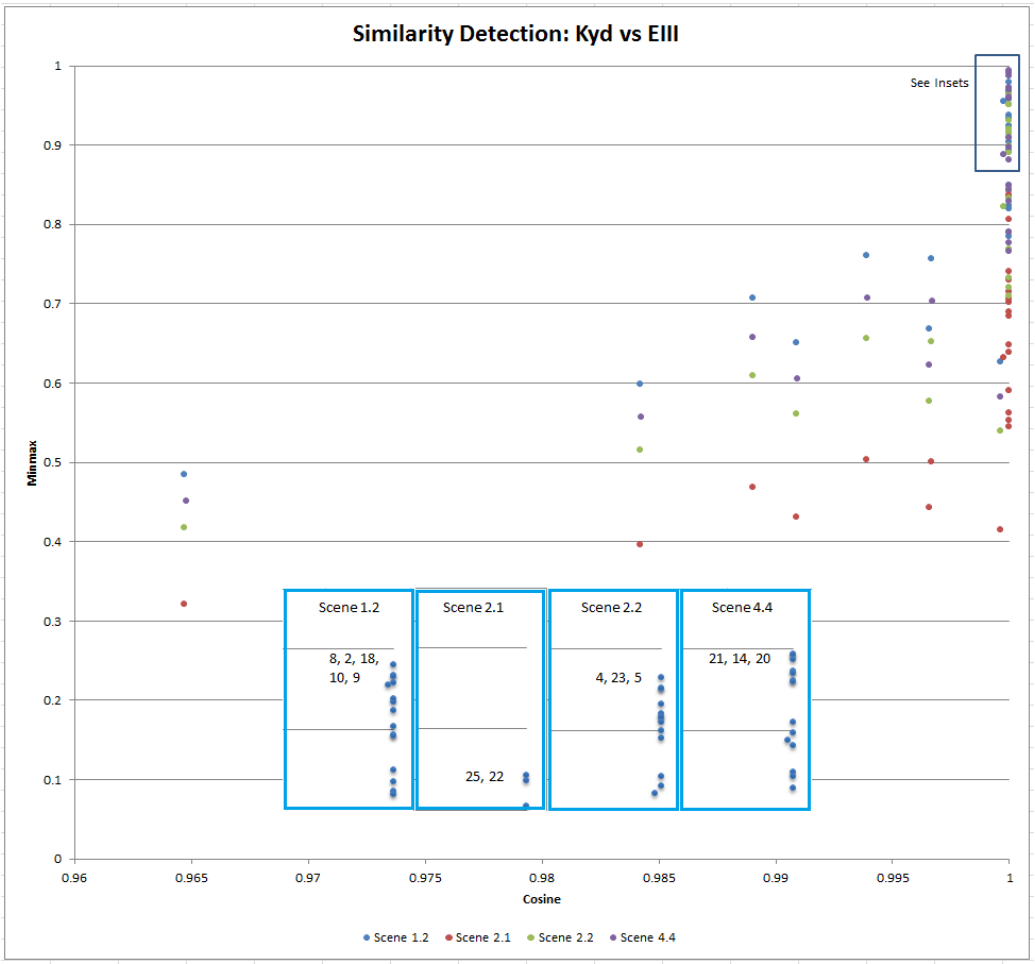


5.3.3 VSM using The Spanish Tragedy Chunks

There are four commonly accepted scenes attributed to William Shakespeare by scholars (clusters 2 – scene 1.2, 3 – scene 2.1, 4 – scene 2.2, and 13- scene 4.4 are red circles in all of the figures) and in each case, we find they sit inside or close to Kyd's clusters. We conduct further VSM analysis to counter the commonly held view that these are Shakespeare's.

We chunk Thomas Kyd's *The Spanish Tragedy* into 25 scenes and conduct pair-wise comparisons of each of them against the four *Edward III* scenes attributed to Shakespeare (100 pair-wise comparisons) using both cosine and minmax similarity detection (see Figure 16). We would expect the Shakespeare scenes to be very dissimilar to Kyd's writing, and the majority of the 25 scenes would be quite far from the top right corner. However, this was not the case. We find 13 of the 25 scene chunks clearly identify with Kyd (~52%), which we believe is a relatively high number given the smaller size comparisons.

Figure 16: Kyd's 25 Spanish Tragedy scenes compared four times (with the four Shakespeare scenes) giving a high number of similarity comparisons.



5.3.4 Seriation

Throughout these different visualization techniques, we see variability with chunks 6, 10, 11, and 18, and the four commonly accepted scenes attributed to William Shakespeare (chunks 2, 3, 4, and 13) identify as Kyd, but many of the techniques have been dependent on an arbitrary visual clustering size. Therefore, to add further reliability to the results, we cluster the data using seriation.

The R seriation package is fed a 9x19 matrix of the data, and using Euclidean distance we seriate the data to minimize the Hamiltonian path length. Results of the six seriation techniques available highlight that Hierarchical Clustering with Optimal Leaf Ordering (OLO) outperforms the Travelling Salesperson technique (path lengths 140.96 vs. 157.52). The order of the 19 chunks is 8 17 15 16 6 10 9 11 18 19 13 14 12 5 1 2 7 4 3 (see Table 5 for more detail). When we compare the distances between each chunk, the ordering sequence is important, but the distance information does not convey much.

Table 5: The Seriation results of the 19 *Edward III* scenes show that Ordinal Leaf Ordering (OLO) technique provides the shortest Hamiltonian path. The 'Shakespeare 5' appears at the beginning distant from the commonly attributed Shakespeare scenes (marked with *).

OLO Order	Chunk	Scene	Author
1	8	3.4	WS
2	17	4.8	WS
3	15	4.6	WS
4	16	4.7	WS
5	6	3.2	WS
6	10	4.1	TK
7	9	4.9	TK
8	11	4.2	TK
9	18	4.5	TK
10	19	5.1	TK
11	13	4.4	TK*
12	14	4.3	TK
13	12	3.5	TK
14	5	3.1	TK
15	1	1.1	TK
16	2	1.2	TK*
17	7	3.3	TK
18	4	2.2	TK*
19	3	2.1	TK*

*scenes traditionally attributed to Shakespeare

To see how stable the results are, we insert noise into the initial 9x19 RPAS-scene matrix and recalculate the Euclidean distances with various amounts of noise (between 1 – 2000). An examination of the scene chunk order after seriation (see Table 6) highlights the susceptibility of chunks 10 and 11 to moderate amounts of noise, and there is some movement of the order of the scene chunks in the middle section with a significant amount of introduced noise. However, we find the Shakespeare chunks (6, 8, 15, 16, 17) do not move, nor do the three chunks commonly attributed to Shakespeare (2, 3, and 4), which remain in a large group with Kyd's work (chunks 1, 2, 3, 4, 5, 7).

Table 6: The different OLO seriation results are showing changes in order when noise is added to the RPAS scene matrix. Shakespeare's work sits on the left (chunks 8, 17, 15, 16, 6), while Kyd's scenes sit on the far right with the commonly accepted Shakespeare scenes (5, 1, 2, 7, 4, 3).

OLO Seriation of 19 Edward III scenes with introduced noise																			Noise
8	17	15	16	6	10	9	11	18	19	13	14	12	5	1	2	7	4	3	0
8	17	15	16	6	10	9	11	18	19	13	14	12	5	1	2	7	4	3	1
8	17	15	16	6	11	9	10	18	19	13	14	12	5	1	2	7	4	3	50
8	17	15	16	6	11	9	10	18	19	13	14	12	5	1	2	7	4	3	100
8	17	15	16	6	10	18	19	13	14	12	11	9	5	1	2	7	4	3	200
8	17	15	16	6	10	18	19	13	12	14	11	9	5	1	2	7	4	3	400
8	17	15	16	6	11	9	10	18	19	14	12	13	7	5	1	2	4	3	800
8	17	15	16	6	10	18	11	12	14	19	13	9	7	5	1	2	4	3	1000
8	17	15	16	6	10	18	11	9	12	14	19	13	5	1	2	7	4	3	2000

5.4 Discussion

When we examine the 19 *Edward III* scenes using **RPAS,** the Visual, Auditory, Haptic, Olfactory, and Gustatory (VAHOG) Sensory (S) results split the data into two distinct groups (a 57/42% split). This split is in line with Brian Vickers' claim of a scene split of about 60/40%, but the data can also be separated by Richness (R) and Personal Pronouns (P) with similar results. The amalgamation of the PtoR, AtoR, and StoR analysis clusters chunks 6, 8, 15, 16, and 17. These chunks have a richer and much wider sensory range with a lesser feminine style (where R > 50 and P > 8), and it highlights the significance of using Richness, Personal Pronouns, and the Sensory VAHOG variables in RPAS. Chunks 10, 11, 12, and 18 stand out but appear to be ambiguous.

By using VSM we can compare *Edward III* to a known work of Thomas Kyd, and we assign chunks 6, 8, 15, 16, and 17 to Shakespeare and the rest to Kyd. Using Shakespeare's *Venus and Adonis*, we again see some further variability in chunk 10 and 11, but overall the chunks are consistent with the previous techniques. These results are reflected in the Imposters Method with VSM using Marlowe's *Hero and Leander*. The only change from Shakespeare's is the order of chunks 6 and 8, and again this reinforces the overall results adding another layer of consistency. Using the imposter method in a study of 42 commonly attributed works of Shakespeare that also included both plays of Thomas Kyd and the Edward III play, the Koppel and Winter (2014) findings suggest that Edward III is more similar to Thomas Kyd's plays than 39 of Shakespeare's.

We also find all four commonly attributed Shakespeare scenes (chunks 2, 3, 4, and 13) consistently fall inside the Kyd clusters or close to them. Using 'new-optics' stylometric measures on Edward III play, Elliot and Valenza (2010) findings suggest that when taken as a group, Shakespeare's authorship of the four scenes commonly attributed to him are unlikely, a view they say is supported in Marina Tarlinskaja (2006) unpublished article.

When we conducted further VSM analysis by using VSM and chunking Thomas Kyd's *The Spanish Tragedy* into 25 scenes shows close similarities to 52% of Kyd's work. The results we believe, given the small size of the chunks, would appear to be a relatively high number of similar works and supports the earlier results that their authorship is probably Kyd's and not Shakespeare's.

We find the arbitrary nature of the clustering size does influence the reporting to a small degree, and while we believe the cluster sizes reasonable, there has been some minor variability with chunks 6 and 18, but more so with chunks 10 and 11. However, using Seriation, it is clear that chunk 6 is part of the 'Shakespeare 5' (sits alongside chunks 8, 15, 16, and 17). Chunks 10, 9, 11, and 18 are the closest chunks to the Shakespeare cluster but are separate from him. By adding a moderate amount of noise to the seriation matrix, we find some variability with chunks 10 and 11. It is possible that they are collaborative scenes containing both the work of Kyd and Shakespeare. However, of the commonly accepted Shakespeare scenes, three of them clustered together (chunks 2, 3, and 4) at the opposite end to the Shakespeare work and no amount of introduced noise moved or separated them from Kyd's work. Only chunk 13 sits away, and while it is six scenes from the 'Shakespeare 5', it is closer to Kyd.

In comparing these results to Vickers' (2014), we find we agree with nine of the scenes that he has suggested are Kyd's, and this analysis suggests that scenes 4.1 (chunk 10) and 4.2 (chunk 11) appear to be Kyd Shakespeare collaborations. We disagree with his analysis of scenes 3.2, 3.4, 4.6, and 4.8 (chunks 6, 8, 15, and 17 from the 'Shakespeare 5' cluster). We also suggest that the four scenes commonly attributed to Shakespeare, scenes 1.2, 2.1, 2.2, and 4.4 (chunks 2, 3, 4, and 13) are written by Kyd, although scene 2.1, and to a lesser extent scene 2.2 is more 'Kyd-like' and away from the main body of the other Kyd scenes (see Elliot & Valenza, 2010 for similarities to these findings). As we show in Table 7, we agree with Vickers' conclusion that the majority of the work *Edward III* was written by Kyd.

Using **RPAS**, we identify subtle characteristics within Shakespeare that identify him separately to Kyd. Shakespeare uses more unique words and less repetition than Kyd, less feminine personal pronouns, and more masculine ones, and overall, he used a wider range of visual descriptions, but draws less on sensory characteristics and emotional experiences, and is vaguer and more general than Kyd.

Table 7: The 19 scenes of the *Edward III* play with the Shakespearian scenes are referenced to the 19 chunks. In the third column, the five commonly attributed Shakespeare scenes are shown against those unknown authored scenes. In column four are the results of Brian Vickers' Kyd trigram scene allocation. Column five shows a summary of the results using RPAS.

Chunk	Scene	Long-Held View	Vickers View	Our View	
1	1.1	U	K	K	
2	1.2	S	S	K	
3	2.1	S	S	K	
4	2.2	S	S	K	
5	3.1	U	K	K	
6	3.2	U	K	S	
7	3.3	U	K	K	
8	3.4	U	K	S	
9	3.5	U	K	K	
10	4.1	U	K	K/S*	Collaboration?
11	4.2	U	K	K/S*	Collaboration
12	4.3	U	K	K	
13	4.4	S	S	K	
14	4.5	U	K	K	
15	4.6	U	K	S	
16	4.7	U	U	S	
17	4.8	U	K	S	
18	4.9	U	U	K	
19	5.1	U	K	K	

5.4.1. A Limitation of the overall approach

A limitation of this overall approach is that the results are dependent on the chunking of the data into 19 scenes. Elliot and Valenza (2010) split scene 2.1 into two parts. If these scenes are not the true delineation between the efforts of two authors, then this would skew the results, but at the end of the day, it is difficult to tell what, if any, a split in the scenes may have been. Here we assume that each scene was written by a single author. However, if this was not true then scene 2.1 would appear stylistically different from both William Shakespeare and Thomas Kyd's other works as a third author. This did not occur, although as we have stated, scene 2.1 split and complete, and to a lesser extent scene 2.2 were more 'Kyd-like' than the other scenes. There were

also two works that were similar to both authors (chunks 10 and 11, or scenes 4.1 and 4.2) and they could well be collaborations. In dealing with over 400 years old text, we suggest the exact details and events that led to this fascinating union of work by Thomas Kyd and William Shakespeare may well be lost within the strands of time.

5.5. Conclusion

In this analysis, the four scenes commonly attributed to Shakespeare identify as Thomas Kyd, but this is not unexpected (see Elliot & Valenza, 2010). However, it seems clear from the analysis that Thomas Kyd wrote the majority of the play and William Shakespeare played a lesser role. On the basis of these findings, the collaborative play, *The Reign of King Edward III,* could well have been written by William Shakespeare and Thomas Kyd.

In examining this multivariate technique, we find the analysis provided a consistent result, and therefore the techniques were resilient. The results of seriation were found to be robust to perturbations in the **RPAS** features and strongly validate the approach to author identification. Significant differentiation was found using RPAS and the neurolinguistics approach of Richness (R), gendered Personal Pronouns (P), Referential Activity power (A), and Sensory modes (S).

5.6 Summary

In this chapter, the use of RPAS equations was found to be effective at identifying the authorship of the Elizabethan poets. The most significant findings for this study were that the different variations of the Vector Space Model (VSM) technique provided consistent results and was supported by the seriation techniques with noise to highlight Kyd likely wrote the four Shakespeare scenes and that Shakespeare and Kyd were the likely authors of the play.

The Passionate Pilgrim

In this chapter, the third and final of the studies into the Elizabethan playwrights' and poets' identity is addressed. The study draws on the known works of William Shakespeare Christopher Marlowe and Sir Walter Raleigh, Richard Barnfield, and Bartholomew Griffin. The study aims to identify the authorship of the 12 unknown-authored poems from the 21 poems within the publication, *The Passionate Pilgrim*, using the same methods as Chapter 5, again addressing the first research question and testing Hypothesis H_1 (Section 1.4) with yet again more complex data.

The results of the PtoR plot, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and the Vector Space Method (VSM) provided consistent clustering results. The unknown poems fell into four of five clusters, where only one poem (12) didn't belong to any of the five authors. **Given these findings, we are able to reject the null hypothesis and say that the stylistic fingerprint of a person's personality – their personal signature – can reveal their 'identity' from their writing style.**

This chapter is taken from a peer-reviewed paper: *Stylometric Techniques for Multiple Author Clustering: Shakespeare's Authorship in The Passionate Pilgrim.* It was published in the *International Journal of Advanced Computer Science and Applications.* (Accepted 14 March 2017 for Vol. 8(3), 1-8).

6.1 Introduction

William Jaggard first printed *The Passionate Pilgrim* in 1598-99, and the authorship of the 21 poems within it was attributed to William Shakespeare (Erne, 2013). However, Bartholomew Griffin's 1596, *Fidessa More Chaste Than Kind,* already contained poem 11 (Devington, 2007). Another, poem 19, appeared anonymously in Anne Cornwallis' 1580 personal notebook alongside works from Sir Philip Sidney, Sir Walter Raleigh, Sir Edward Dyer and Edward de Vere, 17th Earl of Oxford (Woudhuysen, 1996). The list grows, and in 1598, Jaggard's brother John printed Richard Barnfield's, *The Encomion of Lady Pecunia*, containing poems 8 and 11 (Erne, 2013). By 1609, only five had been confirmed as Shakespeare's (poems 1, 2, 3, 5, and 17) having appeared in *The Sonnets*, or his play, *Love's Labour's Lost* (Connor, 2014). Then, England's Helicon also printed a

version of poem 20, attributing it to Christopher Marlowe, although its reply (signed Ignato) was later said to be by Sir Walter Raleigh (Devington, 2007). Jaggard persisted with his claim, and in the 1612 third edition added a number of poems from Thomas Heywood, however, after complaints, Jaggard removed Shakespeare's name from the title (Erne, 2013). By then, the authorship of the 12 remaining anonymous poems lay in doubt, something that has remained for over 400 years.

Modern scholars are divided on the authorship of the remaining anonymous twelve. Chiljan (2012) suggests Jaggard used Shakespeare's name because the majority of the poems were Shakespeare's, including the 12 unidentified poems in *The Passionate Pilgrim* said to be his earlier quality work and never meant for publishing. She also adds there is some doubt surrounding the authorship of the Barnfield and Griffin poems. Bednarz (2007) disputes Shakespeare's authorship, while Elliott and Valenza (1991b) suggest eight, not 12 of the anonymous poems are Shakespeare's. However, Devington (2007) suggest poems 7, 10, 13, 14, 15, 16, and 19 use a similar six-line stanza format to Shakespeare's *Venus and Adonis,* and poems 4, 6, and 9 are *about* Venus and Adonis and have Shakespearian similarities, but Chiljan (2012) says poems 7 and 13 resemble Robert Greene's poems.

It is interesting to note that anonymous poem 12 gets little attention, even though it appears in Thomas Deloney's *The Garland of Goodwill*, and entered into the Stationers Register ledger during 1592-3 (Korp, 2015). When chosen by Jaggard, Deloney was living with an arrest warrant over his head because of his insightful writing during the London riots and in no position to complain (Korp, 2015), but what is strange are the few references in the literature to Deloney as the author until recently. Either way, Jaggard cannot be asked about the true authorship of the 21 poems, and today, the 12 poems, for the most part, remain unidentified.

Stylometric analysis, the quantitative analysis of a text's linguistic features have been extensively used to determine the authorship of the undocumented collaborations of the playwrights from the Elizabethan period, including Shakespeare (Segarra *et al.*, 2017). There appears dissension among leading Shakespearean authorship attribution scholars about an agreed method (Rudman, 2016), but the most successful and robust methods are based on low-level information such as character n-grams or auxiliary words (function word, stop words such as articles and prepositions) frequencies (Stamatatos, 2009). The premier work in evaluating authorship in the 16th to mid-17th centuries includes MacDonald P. Jackson, Brian Vickers, and Hugh Craig and Arthur

Kinney (Segarra *et al.*, 2017). Jackson (2006) uses common low-frequency word phrases, repetition of phrases, collocation, and images to link word groups to other works. Vickers (2011) uses a tri-gram, or n-gram, approach, while Hirch and Craig (2014) use function word frequency and other methods, that includes ones based on word probabilities and the Information Theoretic measure Jensen-Shannon divergence (JSD) and unsupervised graph partitioning clustering algorithms (Arefin *et al.*, 2014). However, there are other techniques used in this period of Shakespearean analysis, including simple function words (Matthews & Merriam, 1993; Merriam & Matthews, 1994) and word adjacency networks (WANs) (Segarra *et al.*, 2017). However, the meaning-extracting method (MEM) from the field of psychology to extract themes from commonly used adjectives and describe a person from their personality, or self is very different (Boyd & Pennebaker, 2015; Chung & Pennebaker, 2008). We offer a new and alternative approach to authorship identification using personality.

6.1.1 An Approach Using RPAS

RPAS is used to create individual stylistic signatures of the 21 *The Passionate Pilgrim* poems and the known works of William Shakespeare, Christopher Marlowe and Sir Walter Raleigh, Richard Barnfield, and Bartholomew Griffin are labelled. Three clustering techniques are then applied to identify the likely authorship of the 12 anonymous poems within *The Passionate Pilgrim*.

6.2 Methodology

The Passionate Pilgrim contained within the complete works of Shakespeare (Farrow, 1993) is pre-processed as per Section 3.1. *The Passionate Pilgrim* is further broken down into chunks that represent each known poem, and a decision made to follow the modern approach by editors (Devington, 2007), and divide poem 14 into two poems (labelled as 14 and 15) with a subsequent renumbering of the remaining poems so that there are twenty-one and not twenty poem chunks (see Table 37 in Appendix A).

The 3,190-word data ends up as an aggregated matrix of 1,032 distinct word types across 21 poems, and the size of each varies between 96 and 377 words (average = 152). Putting this into perspective, they are slightly larger than a Shakespearian sonnet which varies between 91 and 132 words (average = 116).

A 1613 play written after Shakespeare ceased writing is used to provide an independent author perspective and clustering technique. *The Tragedy of Mariam, the*

Fair Queen of Jewry by English poet and dramatist, Elizabeth Cary (Mark, 2014), was published 14 years after *The Passionate Pilgrim*, and stylistically very different to Shakespeare's work and deemed an imposter (see Section 3.11).

A nine-dimensional array is created from the data using RPAS before applying three complementary techniques to reduce any single bias and overlay the results against Richness (R) and Personal Pronoun (P) to determine the possible authorship of the 12 anonymous poems. As a final measure, seriation, an exploratory combinatorial data analysis technique, is used to visualise the nine-dimensional array as a one-dimensional continuum and test the strength of the co-located cluster edges by adding random noise to the data vector.

6.2.1 Complementary Techniques

Principal Component Analysis (PCA) of the 21 poems was conducted (refer to Section 3.14) Linear Discriminant Analysis (LDA) was conducted (see Section 3.15). The Vector Space Method (VSM) technique (Koppel & Winter, 2014; Voorhees, 1998) is used with Elizabeth Cary's, *The Tragedy of Mariam, the Fair Queen of Jewry* as an imposter (Seidman, 2013). Pair-wise comparisons of each of the 21 *Passionate Pilgrim* poems is made against Elizabeth Carey's play (42 pair-wise comparisons) using both cosine and minmax similarity detection, to highlight the clusters that form based on their distance from Cary's play (see VSM Section 3.10 and Imposters method Section 3.11 for more details on the methods).

6.2.2 Seriation

Seriation (refer to Section 3.5) was carried out on the 9x21 matrix consisting of the nine RPAS values for each of the 21 poems of *The Passionate Pilgrim*.

6.3 Analysis

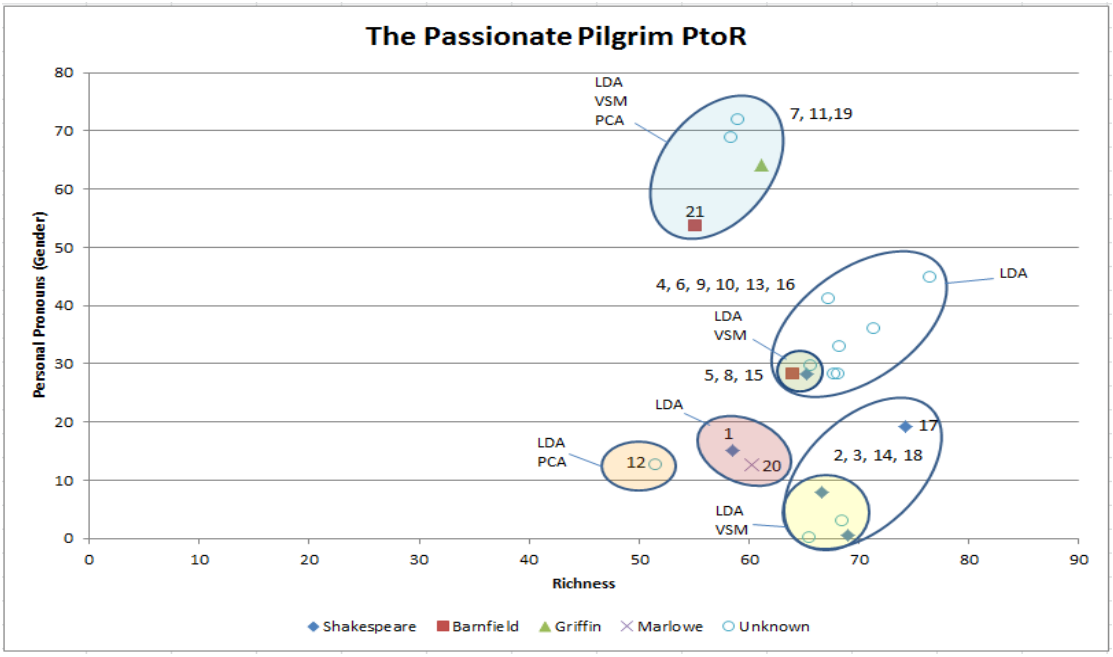
Using **RPAS** Personal Pronouns (P) is plotted against Richness (R) (PtoR) for the 21 *The Passionate Pilgrim* poems (see Figure 17). PtoR discriminates the anonymous poems 14 and 16 with Shakespeare (poems 2 and 3), and they have a low feminine gendered style (P > 10), while all of Shakespeare's known poems have a lower feminine gendered style (P > 30), contrasting this is the group consisting of the cluster with anonymous poems 7 and 19 that are similar in style to Griffin (poem 11) and Barnfield (poem 21) whom all have a higher masculine style (P >50). The Shakespeare (poem 1) and the Marlowe and Walter Raleigh (poem 20) are similar, as are Barnfield (poem 8) and

Shakespeare (poem 5). The anonymous poem 12 (from Deloney) has a low Richness score is separate from the main body of poems.

6.3.1 Principal Component Analysis (PCA)

The findings show that many PCA correlations are in excess of 0.30. A visual indication of the correlation matrix highlights 24 coefficients are around 0.30 or higher and some are as high as 0.77, and Bartlett's test is significant ($p = 0.001$) meaning there is some correlation between variables indicating that PCA is worthwhile. Four components are extracted and account for 81.95% of the variance.

Figure 17: In this *The Passionate Pilgrim* gendered Personal pronouns (P) versus Richness (R) diagram, the overlays of the results of LDA, VSM, and PCA analysis highlight the consistency of other results. A Barnfield / Griffin series of poems can be seen (7, 11, 19, and 21) with greater than 50% gendered personal pronouns. This is supported by LDA, VSM and PCA Analysis. A Shakespeare series of poems can be observed (2, 3, 14, 17, and 18), also supported by LDA and VSM analysis. A Shakespeare / Marlowe / Raleigh series is observed (1 and 20) to have less than 20% gendered personal pronouns supported by LDA analysis. Clearly, Deloney's poem 12 is supported by LDA, and PCA analysis as a standalone work also has the lowest Richness. In the range of 25-50%, gendered personal pronouns are the Shakespeare / Barnfield poems (5, 8, and 15) supported by LDA and VSM analysis, and these alongside the anonymous poems (4, 6, 9) (and 10, 13, 16 supported by LDA analysis). Further, the ellipses are a visual clustering assignment.



In Figure 17, the two common clusters are overlaid. A Barnfield / Griffin group (11 and 21) is found to sit with anonymous poems 7 and 19. While anonymous poem 12 (Thomas Deloney) was close to Shakespeare (1) and Marlowe and Raleigh (20), it is the furthest poem from the Shakespeare cluster on Factor 1 and 2 scale that accounts for ~55% of the variance. Additionally, the results highlight all of the known Shakespeare poems cluster (poems 1, 2, 3, 5, 17 with 6, 14, 15, and 16). Poem 4 is close to Barnfield

(8), and poems 6, 9, 15, and 16 are close to Shakespeare (5). Richness is seen to be very narrow.

6.3.2 Linear Discriminant Analysis (LDA)

Three functions were extracted, and the first two accounted for 99.6% of the variance (1 = 95.9 and 2 = 3.7). The Wilks' Lambda test (Coccia, 2008) of functions 1 through 3 was significant (p=0.009) which highlights that the null hypothesis can be rejected and suggests that all three functions together have a discriminating ability. The second and third functions together are not significant (p=0.190), neither is function 3 on its own (p=0.453). Functions 1-2 and functions 1-3 are plotted to generate six common clustering results (see Figure 17). It is found that the anonymous poems 10 and 13 are again close to Shakespeare (5) and Barnfield (8), as is 15. Anonymous poems 7 and 19 are closer to Griffin (11) this time and further from Barnfield (21). Anonymous poem 12 (Thomas Deloney) is again closest to Shakespeare (1) and Marlowe and Raleigh (20) but stands alone. Poem 14 is again close to Shakespeare (2 and 3).

While poem 18 is also close to Shakespeare (1, 2, and 3), poem 4 is far from all the poems but closest to Griffin (11). Poem 6 is closest to Shakespeare (17). Poem 16 is closest to Shakespeare (5), and poem 9 is in the middle of Shakespeare (5), Barnfield (21) and Griffin (11). Again, there is some consistency with these results, but there seems to be a lack of clarity with poems 4, 6, 9 and 16.

6.3.3 The Vector Space Method (VSM)

Pair-wise comparisons of each of the 21 *Passionate Pilgrim* poems against Elizabeth Carey's play, *The Tragedy of Mariam, the Fair Queen of Jewry* (42 pair-wise comparisons) using both cosine and minmax similarity detection, highlights the clusters that form based on their distance from Cary's play. Figure 17, indicates the three common clustering results. Here, anonymous poems, 7 and 19 are in a cluster with Griffin (11). Anonymous poem 14 is in a cluster with Shakespeare (1, 2, and 3) and Marlowe / Raleigh (20) and poems 12 and 18, and closest to Shakespeare (1), while Deloney's poem 12 and 14 are closest to Shakespeare (2), but furthest away. Anonymous poems 4, 6, 9, 10, 13, 15, and 16 are in a cluster with Shakespeare (5 and 17) and Barnfield (8). In this cluster Barnfield (8) is very close to Shakespeare (5), and poems 10 and 13 have an almost identical score.

Throughout these different analysis techniques, there is a consistency in three to four clusters forming with common poems in them, but many of the techniques have been dependent on a visual clustering size. Therefore, to add further reliability to the results, the data is clustered using seriation to measure cluster distances.

6.3.4 Seriation

The R seriation package is fed a 9x21 matrix of the data, and using Euclidean distance seriation of the data minimizes the Hamiltonian path length. Results of the six seriation techniques available highlight that Hierarchical Clustering with Optimal Leaf Ordering (OLO) outperforms the Travelling Salesperson technique (path lengths 214.63 vs. 228.92). Incorporating the clustering of the OLO dendrogram at a height of 25, the order of the 21 chunks with clusters highlighted is [21 19 7 11] [4 9 6] [5 8 10 13 15 16 17] [20 12 1 3 2 14 18] and it highlights some susceptibility between poems 11-4, 6-5, and 17-20. When the distances between each poem are compared, and either side of poems 11-4 (7-11-4-9), 6-5 (9-6-5-8), and 17-20 (16-17-20-12), the ordering sequence and distance information is important (refer Table 8).

Table 8: Hamiltonian path distances between the 21 *The Passionate Pilgrim* poems. The OLO dendrogram edge clusters that form at a dendrogram height of 25 highlights a consistency in two of the three separation points. In the cluster split at poems 11-4, 7-11 and 4-9 are closer than 11-4 (27.3 versus 11.8 and 9.6). In the cluster split at poems 6-5, 9-6 and 5-8 are closer than 6-5 (10.61 versus 7.7 and 3.4), but in the 17-20 cluster split, while 16-17 and 20-12 are closer than 17-20, the differences between 16-17 and 17-20 are marginal (15.8 and 12.6 versus 16.8).

Poem edges		Path length
21	19	16.60488
19	7	24.69437
7	11	9.561261
11	4	27.27893
4	9	11.78111
9	6	7.683108
6	5	10.61323
5	8	3.444387
8	10	4.88489
10	13	3.22249
13	15	3.455063
15	16	4.449576
16	17	15.8412
17	20	16.75323
20	12	12.6397
12	1	14.13468
1	3	11.68744
3	2	8.28891
2	14	13.00578
14	18	6.162732

Further, when examining the OLO dendrogram edge clusters that form at a dendrogram height of 25 and find consistency in two of the three separation points. In the cluster split at poems 11-4, it can be seen that 7-11 and 4-9 are closer than 11-4 (27.3 versus 11.8 and 9.6). In the cluster split at poems 6-5, 9-6 and 5-8 are closer than 6-5 (10.61 versus 7.7 and 3.4), but in the 17-20 cluster split, while 16-17 and 20-12 are closer than 17-20, the differences between 16-17 and 17-20 are marginal (15.8 and 12.6 versus 16.8).

Table 9: The different OLO seriation results are showing changes in order when noise is added to the RPAS poem matrix. At around noise levels of 500, poems 15 and 16 switch positions, but then revert with further noise. At noise levels 800 and above, the Barnfield – Griffin cluster (7, 11, 19, and 21) move internally within the cluster but no poems leave. At noise levels 800 and higher the Shakespeare – Marlowe cluster (1, 2, 3, 12, 14, 18, 20) move internally. This suggests a high level of stability in the seriation OLO order and OLO clustering results ([21 19 7 11] [4 9 6] [5 8 10 13 15 16 17] [20 12 1 3 2 14 18]).

	Noise							
Order	**0**	**100**	**500**	**800**	**1000**	**2000**	**4000**	**8000**
1	21	21	21	7	7	7	7	7
2	19	19	19	11	11	11	11	11
3	7	7	7	19	19	19	19	19
4	11	11	11	21	21	21	21	21
5	4	4	4	4	4	4	4	9
6	9	9	9	9	9	9	9	6
7	6	6	6	6	6	6	6	4
8	5	5	5	5	5	5	5	5
9	8	8	8	8	8	8	8	8
10	10	10	10	10	10	10	10	10
11	13	13	13	13	13	13	13	13
12	15	15	16	15	15	15	15	15
13	16	16	15	16	16	16	16	16
14	17	17	17	17	17	17	17	17
15	20	20	20	14	20	20	14	14
16	12	12	12	18	12	12	18	18
17	1	1	1	20	1	1	20	20
18	3	3	3	12	3	3	12	12
19	2	2	2	1	2	2	1	1
20	14	14	14	3	14	14	3	3
21	18	18	18	2	18	18	2	2

To see how stable the results are, in particular, the stability of the clusters connected at the poems 17-20 split, noise is inserted into the initial 9x21 RPAS-poem matrix and recalculate Euclidean distances with various amounts of noise (noise 1 – 8000). An examination of the scene chunk order after seriation (refer Table 9) highlights the high level of stability within the seriation and OLO clustering results. The different OLO seriation results are showing changes in order when noise is added to the RPAS poem matrix. At around noise levels of 500, poems 15 and 16 switch positions, but then revert back with further noise. At noise levels 800 and above, the Barnfield – Griffin cluster (7,

11, 19, and 21) move internally within the cluster but no poems leave. At noise levels 800 and higher the Shakespeare – Marlowe cluster (1, 2, 3, 12, 14, 18, and 20) move internally, and at no point does poem 20 moves out of the cluster and join with poem 17.

6.4 Discussion

Overall, the techniques were generally consistent, and seriation was useful because it was able to provide clustering and distance measures that appeared stable even with a relatively high level of introduced noise. Therefore, the basis of these findings lies in a rigorous multivariate approach to analysis and not a single technique. However, one of the biggest concerns is the influence of the publisher. While Jaggard or his associates cannot be discounted from having a hand in adding their own touches to some of these anonymous poems, blending them as it were so they appear as part collaborations, it is an unknown factor. It is known that Jaggard was able to get hold of some of Shakespeare's unpublished work, and both he and his brother John had access to a wide number of Elizabethan works. What cannot be known is how much of this was early unpublished work.

Of the 12 anonymous poems, two are likely Shakespeare's, possibly from his earlier unpublished works (poems 14 and 18 are similar to Shakespeare's poems 2 and 3 and a lesser extent poem 1). However, if they were not earlier Shakespearian poems, then they are from another poet entirely, one that has not been examined. Two other poems (7 and 19) have a blended style similar to Griffin (11) and Barnfield (21), and there is more of Griffin's style (similar to poem 11) in them than Barnfield's, and they are more likely to be Griffin's unpublished work. Again, if they are not an unpublished Griffin poem, then they too are a poet that has not been examined in this paper. Poem (12) has a blended style similar to Shakespeare (1) and Marlowe / Raleigh (20) but consistently shows itself to be different enough to be an independent poet and be the work of Thomas Deloney whose other poems were outside of this analysis.

The remaining seven anonymous poems (4, 6, 9, 10, 13, 15, and 16) are all similar in style to a blended Shakespeare (5 and 17) and Barnfield (8). All of these, as are all of Shakespeare's poems here, have a Richness score over 65%. They all have a Personal Pronoun score below 50%, which would be deemed as a feminine writing style which fits Shakespeare. Poems 4, 6, and 9 are very similar in style to each other and closer to Shakespeare's (5) style than Barnfield (8). Poems 10, and 13 are closer to Barnfield's (8)

style than Shakespeare (5, 17). Poems 15 and 16 have a higher Shakespeare (5) style than Barnfield's (8) and are higher overall from the Shakespeare poems (5 and 17).

This close style of Barnfield's poem (8) to Shakespeare's (5) is an anomaly, and if it were not for the work sitting in the Shakespeare cluster between 5 and 17, then it could be easily said that all the poems (4, 6, 9, 10, 13, 15, and 16) are Shakespeare's. The literature around Richard Barnfield is examined more closely. While Barnfield and Shakespeare were certainly friends (Sauer, 2008) and could have collaborated, these poems are likely to be Shakespeare's because the style of Barnfield's poem (8) is very similar to Shakespeare's poem (5). It has been suggested, that the 1598 version of Barnfield's manuscript obtained by William Jaggard's brother John was of insufficient length (indicated by the sparse printing layout), and William Jaggard provided his brother two poems from the yet unpublished *The Passionate Pilgrim* to extend Barnfield's *Lady Pecunia* publication. In the 1605 reprint of Richard Barnfield's *Lady Pecunia*, the two poems from the 1598 first edition (poems 8 and 21 from *The Passionate Pilgrim*) were not included (Barnfield, 1598; Barnfield, 1605). According to Barnfield (2008), he is said to have claimed authorship of only *one* of the two poems (stylistically likely poem 21). If this is true, then it explains the striking similarities between the Shakespeare and Barnfield poems (5 and 8), and a good indication that Shakespeare wrote both 5 and 8, and therefore poems 4, 6, 9, 10, 13, 15, and 16 are Shakespeare's poems. While it further reinforces Jaggard's approach to borrowing from other author's works, from the analysis it is believed that Shakespeare wrote nine of the twelve anonymous poems (4, 6, 9, 10, 13, 14, 15, 16, and 18) including 1, 2, 3, 5, 17, and 8.

6.5 Conclusion

Given Shakespeare's signature in almost three-quarters of the poems, Jaggard may have adopted shrewd marketing tactics in using Shakespeare's name as the sole author. Indeed, when he expanded the third edition with a collection of nine of Heywood's poems, he did not remove Shakespeare's name from the title, nor did he add Heywood as co-author, but in his collection of assorted verses. Jaggard merely adopted what was a standard convention by publishers in the day (Reid, 2012). The analysis would suggest that the five authors, Barnfield, Deloney, Griffin, Marlowe, and Raleigh were not acknowledged, and several poems may well be collaborative works between Shakespeare and others but this also was common (Thomas, 2000). It is also possible that several poems (7, 14, 18, and 19) are not early work or collaborations, but other

writer's poems not studied here. This failure to acknowledge all author's poems would seem, at least by today's standards, to be an injustice. However, as it can be seen with Jaggard's publication of *The Passionate Pilgrim* and his later publication of Shakespeare's first folio, Jaggard focussed on promoting Shakespeare's work above all others.

6.6 Summary

In this study, the techniques used in study one (a PtoR plot, PCA, and LDA) were combined with the techniques used in study two (VSM and seriation with noise) and applied to a more complicated authorship scenario where there are multiple known and multiple anonymous authors.

Again, the use of RPAS equations was found to be effective at identifying the authorship of the Elizabethan poets. The most significant findings for this study outside of the consistency of the PCA, LDA, VSM, and seriation techniques were that one of the Barnfield poems (poem 8) was identified as Shakespeare's work, and that all anonymous scenes were allocated authorship except one, believed to be written by the poet, Thomas Deloney.

In the next part, the following three chapters look at authorship identity from the perspective of it changing over time.

Part Two: Longitudinal Studies

In part one of this research thesis, the focus was on identifying the authorship of a number of different works using 400-year-old data from the Elizabethan playwrights and poets. In part two, a further Elizabethan study is conducted using William Shakespeare's sonnets to determine if identity changes over time. Then two further studies are conducted using contemporary data from the novels of the authors Iris Murdoch and Phyllis Dorothy (P.D.) James. The first one addresses the research question: Does a person's 'identity' change over time because of life events, such as trauma, depression, and disease, or is it stable? Here we see if we can identity changes over time in a contemporary author's work.

The second one addresses the research question: Can the application of techniques to visualise the critical slowing down phenomena identify changes in a person's moods, or shifts from one state to another, that might indicate a tipping point for self-radicalisation? It is hoped that any observed tipping points observed in writing prior to significant events might mimic changes in the mind of a terrorist prior to them conducting some horrific act.

In study four (Chapter 7) Shakespeare's Dark Lady Sonnets are examined as an effect of writing changes over time. In study five (Chapter 8) Iris Murdoch's and P.D. James' writing is compared using Parts of Speech analysis to show that the data is representative of larger known datasets to highlight known markers for dementia, and in particular, Alzheimer's disease, 10-12 years prior to any formal medical diagnosis. In study six (Chapter 9) the Murdoch and James data is used to examine the effects of RPAS over time, and conduct visualisation techniques and detect the Critical Slowing Down dynamical property to identify any tipping points that might mimic a terrorist's mindset prior to an attack. In these three studies, there are two research questions that are addressed. In the next two studies the second research question is addressed (hypothesis H_2): Does a person's 'identity' change over time because of life events, such as trauma, depression, and disease, or is it stable? In the third study, the third research question is addressed (hypothesis H_3): Can the application of techniques to visualise

the critical slowing down phenomena identify changes in a person's moods, or shifts from one state to another, that might indicate a tipping point for self-radicalisation?

Iris Murdoch was born in Dublin, Ireland, on the 15th July 1919, and she spent much of her life in Oxford and London in England during her writing career. Her success was exemplified when she was made a Commander of the Order of the British Empire in 1976, and a Dame Commander in 1987. Iris Murdoch died in Oxford on the 8th February 1999 (aged 79). P.D. James was born in Oxford, England, on the 3rd August 1920, and she spent much of her life in Oxford and London during her writing career. She was made an Officer of the Order of the British Empire in 1983 and received life peerage as Baroness James of Holland Park in 1991. P.D. James died in Oxford on the 27th November 2014 (aged 94). There are common ties of the same gender, close birth dates, and environmental experiences (both experienced two world wars and the same political and social transitions within England). The novels of P.D. James were used to contrast Iris Murdoch's results because unlike Iris Murdoch who stopped writing once she had been diagnosed with Alzheimer's disease in 1997, P.D. James had not been diagnosed with Alzheimer's disease and she continued to write up until her death.

There are three papers that contribute to in this second part (refer Section 1.6): *Novel Text Analysis for Investigating Personality: Identifying the Dark Lady in Shakespeare's Sonnets; The Impact of Depression and Apathy on Sensory Language;* and *The Stylometric Effects of Aging and Life Events on Identity*. More detail about each one follows in the next three chapters.

The Dark Lady

In this chapter, the last of the four studies into the Elizabethan playwrights and poets are addressed. The study draws on William Shakespeare's collection of sonnets. The study aims to identify if it possible to group different styles within a single author's writing as a precursor to monitoring changes in a person's writing over time and also due to life events that might mimic the stresses a potential terrorist or suicide attacker might face. In this study, 154 sonnets are used to see if it is possible to separate Shakespeare's 'voice' and to highlight the group of *Dark Lady* sonnets. Using **RPAS** described in the methods section (Chapter 3), Seriation with noise is used to cluster the subtle characteristics of Shakespeare's voice.

This fourth study uses data from a single author. In this study, the second research question is addressed (Section 1.3) and hypothesis H_2 is tested (Section 1.4).

In this case, the findings are not clear. The seriation with noise testing was able to cluster the different sonnets into the *Dark Lady, Procreation,* and *Rival Poet* categories. However, we can only conjecture and are unable to determine what life events, such as trauma, depression, and disease, if any may have contributed to these findings. **Given these findings, we are not able to reject the null hypothesis and cannot say that a person's 'identity' changes over time because of life events, such as trauma, depression, and disease.**

This chapter is taken from a peer-reviewed paper: *Novel Text Analysis for Investigating Personality: Identifying the Dark Lady in Shakespeare's Sonnets*. It was published in the *Journal of Quantitative Linguistics*, Vol 24 No 4, 255-272.

7.1 Introduction

Social media has become an important source for information about people and real-world events (Ghajar-Khosravi *et al.*, 2016). Until relatively recently, people who commit crimes have tried to hide their actions and identities, but with the rise of social media, identity is more important, and the public can follow criminal activity as it happens (Ray, 2016). An example of this was the terrorist organization al-Shabaab's use of Twitter to claim responsibility for and live tweet of the Nairobi Westgate Mall attack

in 2013 (Mair, 2016). Terrorists now use social media to spread their message, recruit, and indoctrinate, and Twitter is a key medium (Wright *et al.*, 2016). Equally, lone wolf terrorists make use of social media such as Twitter and Facebook to spread their beliefs and opinions and obtain information to plan an attack (Brynielsson *et al.*, 2013). With the rise of online radicalization, conducting social network analysis (SNA) and identifying the authorship of anonymous blog posts from those who seek to promote terrorism and criminal activity has become paramount for law enforcement agencies, but even when a person's identity is known, conducting content analysis of weblogs and posts to determine their sentiment (mood, anxiety levels, happiness) is equally important (Bermingham, 2009; Kambourakis, 2014; Kaminski, 2013; Yang & Ng, 2007).

We analyse Shakespeare's *Sonnets* because their size falls in between a twitter post and a small weblog post, and as an accepted single-authored work it is reputed to contain several 'voices' (Kambasković-Sawers, 2007). While the construction of a tweet and a small poem structured poem are created in different contexts, both are highly constrained in size with a goal of maximising the power of the message through word choice, and to portray meaning from the symbolic encoding behind language that is greater than the sum of the words.

William Shakespeare's *Sonnets* were first published by Thomas Thorpe and printed by George Eld in 1609 (Duncan-Jones, 1983). It is said they are among the most beautiful and powerful poems in English literature (Rickards, 2014), deeply moving and thought-provoking (Popescu, 2014). The 154 sonnets can be divided into two main sequences. Sonnets 1-126 are addressed to the Fair Youth, an unnamed young man, while sonnets 127-154 address the Dark Lady (Fort, 1933), a promiscuous married woman (Bell, 2008). The presence of the mysterious 'Rival Poet' in sonnets 78-86 falls within the 'Fair Youth' group, and is touted to be Christopher Marlowe, George Chapman, or an amalgam of other contemporaries (Jackson, 2005), while sonnets 1-17 have been singled out as the provocative Procreation sonnets because they encourage the same young man to marry and father children (Crosman, 1990).

What is clear to the reader is that there are three voices: a masculine voice, a feminine voice, and a deliberately disjointed and contradictory speaker at its center to create the effect (Kambasković-Sawers, 2007). The sonnets are split by the speaker's sexual love for the *Dark Lady,* and the spiritual love for the *Fair Youth (*Matz, 2007), but the author's voice is clear in his suspicion of an affair between both his beloveds (Kambasković-Sawers, 2007). It has been suggested that 'Shakespeare the man' can be reconstructed

in the sonnets more completely than from any of his other works (Burnham, 1990), and many believe the sonnets to be autobiographical and cite sonnet one of the Dark Lady sonnets, (sonnet 135) as proof Shakespeare names himself (Stapleton, 1993).

We use **RPAS** (Section 3.4.1), to create individual stylistic signatures of each of Shakespeare's 154 sonnets and from seriation (Section 3.5), we see if we can separate Shakespeare's 'voice' to highlight the *Dark Lady* sonnets. As part of the analysis, we also comment on the *Rival Poet* and the *Procreation* sonnets.

7.2 Methodology

The works of William Shakespeare's *Sonnets* are drawn from the complete works of Shakespeare (Farrow, 1993) and pre-processed (see Section 3.1). The *Sonnets* are further broken down into the 154 individual sonnets and are a maximum of 132 words. We construct a 9-dimensional vector from the results by applying **RPAS** and then provide the seriation package (refer Section 3.5) with 9 RPAS values for the 154 sonnets.

7.3 Analysis

It is clear from the prior construction of a number of stylistic signatures from Elizabethan playwrights' and poets' that the *Sonnets* are the work of Shakespeare, and a stylistic signature can be created from works as small as a sonnet (Chapter 4). What we are uncertain of is if subtle characteristics of the author's 'voice' or personality that reflect mood and tone can be extracted from the short, iambic pentameter form of a sonnet. Given all the sonnets are in the same 14-line format (with the exception of sonnets 99, 126, and 145), rhyme, rhythm, and length are constant (Simonto, 1989). It is, therefore, hoped it would support the identification of an author's subtle characteristics from within these small texts of between 91 – 132 words as it changes over time. Here we conduct two series of tests using seriation in an iterative manner to attempt to maximize the *Dark Lady* group and cluster size and inject random noise into the matrix to examine the strength of the collocated sonnets.

We seriate the data to minimize the Hamiltonian path length. In doing so the seriation package in R analyses the different distances of the 154 sonnets in nine-dimensional space, ordering each sonnet so that we have a single dimension continuum, where each sonnet has the sonnet most similar on either side of it. It does this using the six different techniques (TSP, Chen, ARSA, HC, GW, and OLO) and provides an overall measure of the distance between the two furthest sonnets (at each end of the

continuum), known as the Hamilton path length. In this case, the initial results highlight the Travelling Salesperson (TSP) technique outperforms all of the other five. We then remove each of the RPAS elements one at a time, run seriation again, and see if the path length is reduced without fragmenting the sonnet group clusters, and then return the element. We do this six times, as shown in Table 10 and examine the location of the 28 *Dark Lady* sonnets and the Hamiltonian path length. Each time TSP provides the shortest path. The RPAS configuration that only uses the Richness, Referential Activity Power and Sensory Adjectives (RAS) and the alternate, one that also uses the Personal Pronouns masculine (M) / feminine (F) gender assignment P(G), or expressed fully as elements RP(G)AS, perform equally well, but RAS has a smaller Hamiltonian path size. In Table 10, the masculine (M) and feminine (F) gender element is referred to as G-MF to separate it from the Personal Pronouns score that occurs between 0-1. Importantly, while reducing the number of elements would typically reduce the dimensionality of the data and inherently provide a smaller path length, we find that with our groupings, this has not impacted on our results and smaller dimensions have increased path length reflecting the inherent underlying structure of the data (see RP(P&G)A, otherwise the dimensions have been 8 ± 1).

Table 10: Results of the *Dark Lady* clustering for the different RPAS configurations ordered by path size. Note that RAS and RP(G)AS perform equally well, but RAS has a smaller Hamiltonian path size. Path size is weakly correlated with dimensionality. If we can assume a single-authored work, then the Masculine / Feminine Gender aspects of the Personal Pronouns (A) element A(G) is not required, and RAS is superior.

Grouping	Removed	Clusters	Cluster sizes	Path size
Dark Lady Sonnets (ideal)		1	28	1
RP(P&G)AS	-	2	25 3	1229
RP(P&G)A	S	4	24 2 1 1	1183
RP(P)AS	Gender (G-MF)	2	25 3	1158
RP(P&G)S	A	3	25 2 1	1060
P(P&G)AS	R	2	26 2	926
RP(G)AS	P Score	1	28	809
RAS	P (Score & G-MF)	1	28	760

We also examine the nine *Rival Poet* sonnets (78-98), and the 17 *Procreation* sonnets (1-17) to see how they cluster with the different configurations (see Table 11). We find that RP(G)AS outperforms all other RPAS configurations by clustering both groups, with RAS also forming a group of 16 *Rival Poet* sonnets, but it does not do as well with the nine *Procreation* sonnets (a cluster of 5 and 4), although RAS has a shorter Hamiltonian path (760 vs. 809). In performing further iterative seriation, we remove the R, A, and S elements from the RAS configuration and find the clusters worsen (fragment further or lose structure).

To see how stable the results are, we insert noise into the initial 9x154 RPAS-sonnet matrix and recalculate Euclidean distances using various amounts of noise (between 1 – 4000). An examination of the sonnet order after seriation (see Table 12) highlights the susceptibility of sonnets 127, 128, and 132 to moderate amounts of noise and sonnets 124 and 126 are introduced into the *Dark Lady* cluster. The *Rival Poet* sonnets have a minimal susceptibility to the introduced noise, with movement occurring in sonnet 12 only, and we find the *Procreation* sonnets are sensitive to small amounts of noise with the cluster constantly forming into different group sizes. What is interesting, are that sonnets 18 and 19 tend to shift with the procreation sonnets, even under small amounts of introduced noise.

Table 11: Results of the *Rival Poet* and *Persuasion* sonnet clustering for different RAS configurations. Note that RP(G)AS performs better than RAS, but we prefer to use RAS because it has a smaller Hamiltonian path. With the exception of RP(P&G)A = 4, all other groupings are 8 ±1.

Grouping	Removed	*Rival Poet* Clusters	Sizes	*Procreation* Clusters	Sizes
RP(P&G)AS	-	3	10 6 1	2	8 1
RP(P&G)A	S	3	13 3 1	2	8 1
RP(P&G)S	A	2	14 3	2	8 1
P(P&G)AS	R	2	14 3	2	8 1
RP(P)AS	G-MF	3	10 6 1	2	8 1
RAS	P (Score & G-MF)	1	16	2	5 4
RP(G)AS	P Score	1	16	1	9

Table 12: The different TSP seriation results showing changes in order when noise is added to the RAS sonnet matrix. Shakespeare's *Dark Lady* Sonnets move position, but the cluster only splits because of the interaction of Sonnet 126, which is the last of the Young Man/Fair Youth sonnets and has a different structure from all of the other sonnets. The Rival Poet sonnets are stable with minor fragmentation of sonnet12 around noise levels of 400, while the Procreation sonnets are quite susceptible to noise but follow sonnets 18 and 19.

	Noise									
Order	0,	1,	50,	100,	200,	400,	800,	1000,	2000,	4000,
1	53	120	53	140	65	72	38	120	65	77
2	49	117	49	139	66	71	35	117	66	70
3	51	116	51	141	60	74	34	114	60	69
4	57	114	57	145	59	75	33	113	59	68
5	58	113	58	144	56	76	32	112	56	66
6	61	112	61	147	54	79	30	109	54	65
7	63	109	63	148	55	81	29	105	55	60
8	62	105	62	150	52	83	26	106	52	59
9	71	103	71	149	50	84	24	102	50	56
10	72	98	72	152	48	88	23	103	48	54
11	73	99	73	154	45	90	22	98	45	55
12	75	100	75	153	44	92	21	91	44	52
13	76	97	76	151	38	91	20	92	41	50
14	79	95	79	146	35	98	18	90	38	48

Noise										
Order	0,	1,	50,	100,	200,	400,	800,	1000,	2000,	4000,
15	78	89	78	143	34	102	19	88	35	45
16	80	87	80	142	33	103	16	84	34	44
17	82	85	82	138	32	106	15	83	33	38
18	86	86	86	137	30	105	14	81	32	35
19	85	82	85	135	27	109	13	79	30	34
20	87	80	87	136	25	112	10	78	25	33
21	89	78	89	134	19	113	6	77	27	32
22	94	73	94	133	16	114	3	70	28	30
23	95	67	95	130	15	120	2	69	21	27
24	93	64	93	131	14	117	1	68	20	25
25	96	63	96	129	13	121	4	64	19	16
26	101	61	101	124	10	123	5	67	18	15
27	100	58	100	127	6	124	7	66	17	14
28	97	57	97	132	3	127	8	65	15	13
29	104	62	104	128	4	129	9	60	16	10
30	106	53	106	126	2	131	11	59	7	6
31	108	42	108	125	1	130	12	56	5	3
32	107	47	107	122	5	133	17	54	1	2
33	110	46	110	118	7	134	25	55	2	1
34	111	43	111	123	8	136	27	51	4	4
35	115	40	115	121	11	135	28	49	8	5
36	118	41	118	119	9	137	31	52	9	7
37	122	36	122	115	12	138	36	50	3	8
38	125	39	125	116	18	142	37	48	6	9
39	126	37	126	111	20	149	39	45	10	11
40	128	31	128	110	17	150	40	44	11	12
41	127	29	127	107	21	152	41	38	12	17
42	132	26	132	108	22	148	44	35	14	18
43	139	27	139	104	23	153	45	34	13	19
44	141	28	141	100	24	154	48	33	22	20
45	143	24	143	101	26	151	50	32	23	21
46	144	23	144	99	28	146	49	30	24	22
47	145	22	145	97	29	145	51	25	26	23
48	140	21	140	95	31	147	52	20	29	24
49	146	18	146	94	36	144	54	19	31	28
50	147	19	147	96	37	143	55	16	36	26
51	151	20	151	93	39	141	59	15	37	29
52	153	17	153	89	41	139	60	17	39	31
53	154	12	154	87	40	140	56	14	40	36
54	152	11	152	85	42	132	61	11	42	37
55	150	7	150	86	47	128	63	8	47	39
56	149	5	149	82	46	126	64	9	46	42
57	148	4	148	80	43	125	68	7	43	47
58	142	1	142	73	49	122	69	5	51	46

Noise										
Order	0,	1,	50,	100,	200,	400,	800,	1000,	2000,	4000,
59	138	2	138	78	51	118	70	1	49	43
60	137	6	137	77	57	119	65	2	53	40
61	135	3	135	70	58	116	66	4	57	41
62	136	8	136	69	53	115	67	3	58	49
63	134	9	134	68	62	111	73	6	62	51
64	133	10	133	66	61	110	78	10	61	53
65	131	13	131	65	63	108	80	13	63	57
66	130	14	130	60	64	107	82	12	64	58
67	129	15	129	59	67	104	86	18	67	62
68	124	16	124	56	68	97	85	21	69	61
69	123	25	123	55	70	95	87	22	70	63
70	121	30	121	54	69	100	89	24	68	64
71	119	32	119	52	73	101	94	23	73	67
72	116	33	116	50	74	99	95	26	76	73
73	117	34	117	48	71	93	97	28	75	78
74	120	35	120	45	72	96	100	27	72	80
75	114	38	114	44	75	94	101	29	71	82
76	113	45	113	41	76	89	102	31	74	86
77	112	44	112	38	83	87	103	36	77	85
78	109	48	109	35	84	85	106	37	78	87
79	105	50	105	33	80	86	105	39	79	89
80	103	49	103	34	78	82	109	41	81	94
81	102	51	102	32	82	80	112	40	88	95
82	99	52	99	30	86	78	113	43	91	97
83	98	55	98	28	85	73	114	46	90	100
84	91	54	91	27	87	67	117	47	92	104
85	92	56	92	25	89	66	120	42	93	107
86	90	59	90	20	94	65	130	53	96	108
87	88	60	88	17	95	60	138	57	101	110
88	83	65	83	12	97	59	134	58	99	111
89	84	66	84	9	100	56	136	61	98	115
90	81	68	81	11	104	55	133	63	103	116
91	74	69	74	8	106	54	135	62	102	119
92	77	70	77	7	111	52	137	72	106	118
93	70	77	70	5	110	51	142	71	105	122
94	69	74	69	4	107	49	143	74	109	125
95	68	72	68	1	108	50	144	75	112	126
96	64	71	64	2	116	48	145	76	113	128
97	67	75	67	3	115	45	147	73	114	127
98	66	76	66	6	118	44	149	80	117	124
99	65	79	65	10	119	41	150	82	120	129
100	60	81	60	13	121	38	152	86	121	132
101	59	83	59	14	123	35	148	85	123	131
102	56	84	56	15	124	34	153	87	124	133

Noise										
Order	0,	1,	50,	100,	200,	400,	800,	1000,	2000,	4000,
103	54	88	54	16	127	33	154	89	127	135
104	55	91	55	19	126	32	151	94	129	137
105	52	90	52	18	125	30	146	93	131	142
106	50	92	50	21	122	25	140	96	130	148
107	48	93	48	22	128	20	141	101	133	152
108	45	94	45	23	132	19	139	99	135	149
109	44	96	44	24	140	16	132	100	136	150
110	38	101	38	26	139	15	128	95	134	154
111	35	102	35	29	141	7	131	97	138	153
112	34	106	34	31	143	5	129	104	137	151
113	33	104	33	36	144	4	127	107	142	147
114	32	107	32	37	147	1	124	108	143	146
115	30	108	30	39	145	2	121	110	148	140
116	27	110	27	40	146	3	123	111	152	145
117	28	111	28	43	151	6	126	116	150	144
118	26	115	26	46	154	9	125	119	149	141
119	24	118	24	47	153	8	122	121	153	139
120	22	119	22	42	152	11	118	123	154	143
121	23	121	23	53	149	10	119	115	151	138
122	21	124	21	51	150	13	116	118	146	136
123	20	123	20	49	148	14	115	122	147	134
124	17	122	17	57	142	12	111	125	144	130
125	19	125	19	58	137	18	110	126	145	120
126	18	126	18	62	138	17	107	128	141	123
127	12	128	12	63	135	21	108	132	139	121
128	8	127	8	61	133	22	104	127	140	117
129	9	129	9	64	136	24	98	124	132	114
130	10	131	10	67	134	23	99	129	128	113
131	6	132	6	71	130	26	96	131	126	112
132	3	130	3	72	131	28	93	130	125	109
133	4	138	4	74	129	27	92	134	122	105
134	2	134	2	75	120	29	90	136	119	106
135	1	136	1	76	117	31	91	133	118	102
136	5	133	5	79	114	36	88	135	115	103
137	7	135	7	81	113	37	84	137	116	98
138	11	137	11	83	112	39	83	138	111	99
139	13	142	13	84	109	40	81	142	110	101
140	14	143	14	88	105	42	79	143	108	96
141	15	149	15	90	103	43	77	141	107	93
142	16	150	16	92	102	46	74	139	104	92
143	25	152	25	91	98	47	76	140	100	90
144	29	148	29	98	99	53	75	146	97	91
145	31	153	31	102	101	57	72	145	95	88
146	36	154	36	103	96	58	71	144	94	84

Noise										
Order	0,	1,	50,	100,	200,	400,	800,	1000,	2000,	4000,
147	37	151	37	106	93	62	62	147	89	83
148	41	146	41	105	92	63	58	151	87	81
149	40	147	40	109	91	61	57	154	84	79
150	39	144	39	112	90	64	53	153	83	76
151	42	145	42	113	88	69	47	148	80	75
152	47	141	47	114	81	68	46	149	85	74
153	46	139	46	117	79	70	43	150	86	71
154	43	140	43	120	77	77	42	152	82	72

We examine the Hamiltonian path distances between the *Dark Lady* sonnets taking note of their distances to adjoining sonnets (see Table 13) to see why sonnets 127, 128, and 132 are susceptible to moderate amounts of noise. We find the Hamiltonian path distances of the two *Dark Lady* edge sonnets (127 and 128) are closer to the non-Dark Lady ones (124 and 126) between the 26 Dark Lady sonnets, including the edge sonnets 124, 125, 126. Some of the results are from the noise jitter we introduced, but this is likely due to the unusual 'non-iambic pentameter' structure of sonnet 126. Sonnet 126 is different in structure from the other sonnets, not only because it is smaller, or a short stanza, but it is also a concluding one and a juncture between the end of the Fair Youth sonnets and the beginning of The Dark Lady ones. It has 12 lines consisting of six rhymed couplets instead of Shakespeare's normal 14 lines of five rhymed couplets, so there are fewer linked themes, but each has more content.

Table 13: Hamiltonian path distances between the 26 Dark Lady sonnets, including the edge sonnets 124, 125,126 where 124 and 126 are the Fair Youth sonnets. We find 127 closer to 124 (Fair Youth) and 128 closer to 126 (Fair Youth), most likely attributed to the unusual 'non-iambic pentameter structure of sonnet 126.

Co-located Sonnets		Distance
124	127	3.645803
127	129	6.165869
129	131	6.675812
131	130	7.896462
130	138	8.454266
138	134	6.252812
134	136	5.208263
136	133	5.820263
133	135	3.070293
135	137	3.467677
137	142	5.394506
142	143	3.505065
143	148	8.592909
148	152	7.145475
152	149	5.761041

Co-located Sonnets		Distance
149	150	1.50537
150	153	8.574628
153	154	2.235739
154	151	3.26277
151	146	6.238941
146	147	5.676883
147	144	3.813679
144	145	4.282785
145	141	4.994587
141	139	3.27575
139	140	6.189313
140	132	9.30889
132	128	5.420591
128	126	3.737001
126	125	2.867308

7.4 Discussion

When we examine the 154 sonnets using 10 different configurations of RPAS, we find we can optimize the algorithm and cluster the 28 *Dark Lady* sonnets using two different configurations. Using RAS, the number of unique words used by an author (Richness), function words, or word particles used to identify clinical depression (Referential Activity Power), and the way people interpret images and concepts through their visual, auditory, haptic, olfactory, and gustatory senses (VAHOG sensory elements) provides the shortest Hamiltonian path score using the Travelling Salesperson (TSP) seriation method. This can also be achieved with an RP(G)AS configuration that includes the pronouns closely aligned to gender and self (Personal Pronouns) indicating a person's writing style or gender (G) as masculine or feminine, but this is achieved with a longer TSP Hamiltonian path score (809 vs.760).

We also find RAS groups the 16 *Rival Poet* sonnets, and forms two small clusters of the nine *Procreation* sonnets, while RP(G)AS outperforms all other RPAS configurations by clustering both groups (see Table 14). We do find that sonnet 18 19 follow the procreation sonnets and thematically are tied through the mention of time, and it could be that they are part of the procreation group and if we included them much of the procreation fragmentation disappears. Pilla (2012) suggests that sonnet 19 is sometimes seen as the last of the Sonnets group.

By introducing noise, we see the influence of sonnet 126 (structurally different from most other sonnets) on the *Dark Lady* sonnets, but are able to show the strength of the *Dark Lady* sonnets' 'voice'. This is true of the *Rival Poet* sonnets also, but not of the *Procreation* sonnets whose structure is more tenuous and breaks into different sized groups with a small amount of noise. Taking the noise testing into account highlights the similarity of the RAS and RP(G)AS clustering results, but RAS still has a smaller Hamiltonian path score, and if we assume we know the identity of the author then the P(G) element is not required and we can state that the RAS configuration is superior.

7.5 Conclusion

RPAS provides a clear ordering of the Shakespeare sonnets, differentiating the Dark Lady, Procreation and Rival Poet categories. No other technique has so far succeeded in doing this. Since the smallest sonnet here is 91 words, the technique could be applied to short blog posts and thus could be used in a range of problems where authorship or author-state-of-mind is sought.

Table 14: Iterative Seriation of the 9-dimensional RPAS vector, from left to right, showing the clustering of the 28 *Dark Lady* sonnets (127-154), the 9 *Rival Poet* sonnets (78-86), and the 16 Procreation sonnets (1-17). Note that the RAS configuration performed best for single-authored works. It had the smallest Hamiltonian path distance, but the RP(G)AS configuration clustered all three groups.

RP(P&G)AS												
41	11	9	3	6	4	1	5	7	8	12	18	25
39	43	53	58	55	56	59	67	68	69	70	77	73
84	95	94	93	96	98	106	104	126	143	138	129	123
116	107	108	110	111	115	119	118	122	125	128	132	140
139	141	145	144	147	146	151	154	153	148	152	150	149
142	137	135	136	134	133	130	131	127	124	121	120	117
114	113	112	109	105	103	102	101	100	97	99	92	90
91	88	89	87	83	85	86	82	80	78	79	81	76
75	74	72	71	62	63	61	64	66	65	60	57	54
52	50	48	49	51	46	47	42	40	44	45	38	35
34	33	32	37	36	31	29	30	27	28	26	24	23
22	21	20	17	19	16	15	14	13	10	2		
No Sensory (S) VAHOG												
152	150	149	148	153	154	151	146	147	144	145	140	141
139	142	137	135	133	136	134	130	131	132	127	124	128
125	122	118	119	121	120	117	114	113	112	109	105	102
103	110	111	115	108	107	97	100	101	99	92	90	91
88	89	87	85	86	82	80	83	81	79	78	76	75
74	72	71	66	65	60	64	61	63	62	57	54	52
49	51	46	47	42	40	50	48	45	44	38	35	34
33	32	30	36	37	31	29	27	28	26	24	23	22
21	20	17	19	16	15	14	13	10	2	6	4	1
5	7	8	12	18	25	3	9	11	41	39	43	53
58	55	56	59	67	73	69	68	70	77	84	93	96
94	95	98	104	106	116	123	129	126	138	143		
No Referential Activity Power (A)												
39	43	53	58	55	56	59	67	68	69	70	73	77
84	95	94	93	96	98	104	106	116	123	129	120	112
109	113	114	117	121	124	127	130	131	133	134	136	135
137	142	139	141	144	148	152	149	150	153	154	151	147
146	145	140	132	128	125	122	118	119	115	111	110	108
107	97	100	99	102	103	105	101	92	91	90	88	89
87	85	86	82	80	78	83	81	79	76	75	74	72
71	66	65	60	64	63	62	61	57	54	52	50	48
45	44	49	51	47	46	42	40	36	37	38	33	35
34	32	30	27	29	31	28	26	24	23	22	21	20
19	16	15	17	14	13	10	2	6	4	1	5	7
8	12	18	25	3	9	11	41	126	138	143		

Continued on next page

No Richness (R)												
53	55	56	59	58	67	73	68	69	70	77	84	93
95	94	96	98	104	106	116	123	129	140	142	146	148
153	154	152	151	150	149	147	145	144	141	139	137	136
135	132	133	134	130	131	128	127	125	124	122	121	119
118	120	117	114	112	113	115	111	110	109	108	107	105
103	102	101	100	99	97	92	91	90	89	87	88	86
85	83	82	79	81	80	78	74	76	75	72	71	62
63	64	66	65	61	60	57	54	52	51	50	49	48
47	46	44	45	42	40	38	35	36	37	34	33	32
31	29	30	28	27	26	24	23	22	21	20	19	17
16	15	14	13	10	2	1	4	6	5	8	7	12
18	25	39	43	3	9	11	41	126	138	143		
No Personal Pronoun (P)												
109	112	113	114	117	120	116	119	121	123	124	129	131
130	133	134	136	135	137	138	139	143	142	149	150	152
148	153	154	151	146	147	144	145	141	140	132	127	128
126	125	122	118	115	111	110	108	107	104	106	105	103
102	98	99	101	100	97	95	94	96	93	92	90	91
88	89	87	85	86	82	80	78	83	84	81	79	76
75	74	72	71	73	67	64	61	63	62	58	57	51
49	53	42	47	46	43	40	39	37	36	31	29	27
28	26	24	23	22	21	18	19	16	15	14	13	10
6	3	2	1	4	5	7	8	9	11	12	17	20
25	30	32	33	34	35	38	41	44	45	48	50	52
55	54	56	59	60	66	65	68	69	70	77		
No Personal Pronouns Gender (G)												
41	11	9	3	6	4	1	5	7	8	12	18	25
39	43	53	58	55	56	59	67	68	69	70	77	73
84	95	94	93	96	98	106	104	126	143	138	129	123
116	107	108	110	111	115	119	118	122	125	128	132	140
139	141	145	144	147	146	151	154	153	148	152	150	149
142	137	135	136	134	133	130	131	127	124	121	120	117
114	113	112	109	105	103	102	101	100	97	99	92	90
91	88	89	87	83	85	86	82	80	78	79	81	76
75	74	72	71	62	63	61	64	66	65	60	57	54
52	50	48	49	51	46	47	42	40	44	45	38	35
34	33	32	37	36	31	29	30	27	28	26	24	23
22	21	20	17	19	16	15	14	13	10	2		

Continued on next page

No Personal Pronouns (A) including no Gender (G)												
38	35	34	33	32	30	29	26	24	23	22	21	18
19	16	15	14	13	10	6	3	4	2	1	5	7
8	9	11	12	17	20	25	27	28	31	36	37	39
40	41	44	45	48	50	52	55	54	56	59	60	66
65	68	69	70	77	74	71	72	75	76	79	81	83
84	88	91	90	92	93	94	96	101	100	99	98	102
103	105	109	106	104	108	110	111	114	113	112	120	117
116	119	121	123	124	127	129	131	130	138	134	136	133
135	137	142	143	148	152	149	150	153	154	151	146	147
144	145	141	139	140	132	128	126	125	122	118	115	107
97	95	89	87	85	86	82	80	78	73	67	64	61
63	62	58	57	51	49	53	42	47	46	43		

7.6 Summary

In this study, the seriation with noise technique was applied to a single-authored work. The most significant findings were that when using seriation, the subtle characteristics of a person performed best with an RP(G)AS configuration, another word, using gender and masculine or feminine over the actual personal pronouns score. However, if the gender of the author was known, then those aspects could be discarded and a configuration of RAS also was just as effective in examining the subtle characteristics of a person's personality.

In the next chapter, a comparative longitudinal study of the contemporary novelists, Iris Murdoch and P.D. James is conducted to determine if a person's 'identity' change over time because of life events, such as trauma, depression, and disease.

Language Markers Using POS Analysis

This chapter is the second of the three studies looking at the changes in an individual's writing style over time, and the first of the two studies using the novels of Iris Murdoch and P.D. James. In the previous study (Chapter 7 or study four), the focus was on examining subtle changes within a single author's work that had reported to be written over a long period of time. However, Shakespeare's collection of sonnets is over 400 years old. In study five, the works of contemporary authors are collected to examine the effects of Alzheimer's disease on RPAS over time. The study aims to compare Iris Murdoch's and P.D. James' writing using Parts of Speech analysis to show that the data is representative of larger known datasets to highlight known markers for dementia, and in particular, Alzheimer's disease. A feature of our analysis is identified 10-12 years prior to any formal medical diagnosis.

In this study, the second research question is addressed again (Section 1.3), but this time using contemporary data to test hypothesis H_2 (Section 1.4).

The Parts of Speech analysis, Mann-Whitney U-Testing, and Principal Component Analysis identified both normal aging and changes over time because of life events. **Given these findings, we are able to reject the null hypothesis and say that a person's 'identity' changes over time because of life events, such as trauma, depression, and disease.**

This chapter is taken from a peer-reviewed paper: *The Impact of Depression and Apathy on Sensory Language*. It was published in the *Open Journal of Modern Linguistics* (Volume 7, 8-32, 2017).

8.1 Introduction

The debate over the notion that language strongly influences thought (Whorf, 1997) is met equally by those who argue that language does not influence it, but historically, language was thought to be tied to an ability to form thoughts (Wicklund, Johnson, & Weintraub, 2004). Some believe that language and thought are combined to modify language and thought further (Ammar & Gohar Ayaz, 2016). Others suggest human

language is an instrument of thought and communicates the attributes of human culture (Lieberman, 2016), or that language allows us to share the knowledge and experiences of others to increase our mental resources (Corballis, 2016). Through this embodied cognition, our concepts are grounded in our sensory and motor systems to develop new abstract representations (Jamrozik *et al.*, 2016). We would argue that the way we think comes through clearly in the multimodal sensory elements of our language and that these aspects of language (such as through sound and visual body language cues) impact on thought, but that disease impacts these.

While thought strongly influences language, depression, and apathy severely impact thinking, they can occur without dementia. There is a close link with depression in dementia, and apathy and depression are the most frequent neuropsychiatric symptoms in one type of dementia, Alzheimer's disease (AD) (Robert, Bremond, & David, 2016). In a different kind of language-based dementia, known as Primary Progressive Aphasia (PPA), patients who could not find the right words to express their thoughts, could still demonstrate they could think clearly (Fedorenko & Varley, 2016; Wicklund, Johnson, & Weintraub's, 2004). While language in PPA is a prominent dysfunction for the first two years of the disease, Alzheimer's disease comes to medical attention because of forgetfulness, usually accompanied by apathy, but not from language dysfunction (Mesulam, 2003). Personal identity, self, persists far into the end stage of the disease (Sabat, & Harré, 1992), while apathy is characterized by reduced motivation, social disinterest, and emotional blunting in the absence of mood-related changes (Chau *et al*., 2016).

If writing is the ability to think and put language on paper or some other visual medium, then the impact of depression and apathy on thought and language might be measurable, but this concept is not new. We draw on a number of recent articles on Iris Murdoch and Alzheimer's disease progression where it is stressed that the author's works had little to no editorial changes (Garrard *et al.*, 2005; Hirst, & Wei Feng, 2012; Lancashire, & Hirst, 2009; Le *et al.*, 2011; Pakhomov *et al.*, 2011; van Velzen, Nanetti, & de Deyn, 2014).

To test the hypothesis that thought and language are impacted by depression and apathy and revealed in a person's writing style 12 years before a formal diagnosis of Alzheimer's disease presents, we draw on earlier studies of AD. We use the novels of Iris Murdoch and P.D. James, however, we use a larger, more complete set than previously used by Garrard *et al.* (2005) and Le *et al.* (2011). We use broader Parts of

Speech analysis techniques, and we also use a new analytical technique from sensory adjectives, to determine what can be seen in language from the impact of depression and apathy in the early onset of Alzheimer's disease.

8.2 Methodology

In this section, we discuss the existing studies conducted into identifying Alzheimer's disease through writing. We draw on two of the RPAS variables, Richness (R) which have been described already (Section 3.4.1). We also draw on Sensory Adjectives (S) (Section 3.4.1), but there is a more detailled description provided here. We also describe what is done with the data in the pre-processing stage so that we can visualise markers for depression, apathy, and Alzheimer's disease in language.

8.2.1 Existing Alzheimer's disease (AD) markers

The idea that cognitive decline in Alzheimer's disease is visible in writing appeared in Snowdon *et al.*'s (1996) findings of a longitudinal study of 678 Catholic sisters. Known as the Nun Study, researchers were able to correlate post-mortem markers for AD in the sister's brains to the density of ideas (from Kintsch & Keenan, 1973 and Turner & Green, 1977) expressed in sentences using Parts of Speech (POS) Tag analysis. Idea density uses elements of language, verbs, adjectives, adverbs, prepositions, and conjunctions divided by the number of words to create a measure of cognitive ability (Brown *et al.*, 2008). Garrard *et al.* (2005) were also instrumental in highlighting Alzheimer's disease through changes in writing and used a different approach which included some other elements of language (nouns, verbs, adverbs and adjectives and function words, e.g., conjunctions, and pronouns) to create word lists. Garrard *et al.* (2005) highlighted significant lexical differences between in Iris Murdoch's early writing and her final novel using fully-parsed texts. Extending these findings, Le *et al.* (2011) conducted a large-scale longitudinal study of Iris Murdoch using 20 fully parsed texts and included P.D. James as a non-Alzheimer and healthy control with no known cognitive decline using 15 of her fully parsed novels. Using the Type Token Ratio to measure lexical richness, Le *et al.* (2011) identified a decline in Iris Murdoch's lexical richness over time and a dip in the data in the middle of her writing career. As Ahmed *et al.* (2013) and Ferguson *et al.* (2014) point out; the study of the subtle language changes over the lifespan of well-known writers (Lancashire, 2010), including Iris Murdoch and Agatha Christie (eg Garrard *et al.*, 2005; van Velzen & Garrard, 2008; Lancashire & Hirst, 2009; Le, 2010; Le *et al.*, 2011) and political figures (Garrard, 2009)

has highlighted that Alzheimer's disease may be apparent years or even decades anyone becomes aware of any symptoms of cognitive deterioration. AD is apparent through lexical repetition and is marked by smaller, higher frequency vocabulary and lower use of Function Words over Content Words (Bird *et al.*, 2000; Garrard *et al.*, 2005). A recent study suggests that Alzheimer's disease can be seen in people's writing 10-12 years before the disease is diagnosed (Rajan *et al.*, 2015).

8.2.2 Sensory Adjectives (S)

While apathy is characterised by reduced motivation, social disinterest, and emotional blunting in the absence of mood-related changes, it has been associated with low norepinephrine levels in the brain (Chau *et al.*, 2016). In the following paragraphs, we describe the link between apathy and depression in people to different levels of norepinephrine in the brain, and how it might be apparent in sensory processing and impact the sensory language of Adjectives.

Many mental disorders have also been associated with alterations of neurotransmitters in the brain (Heilman, Nadeau, & Beversdorf, 2003; Nonen *et al.*, 2016; Sun, Hunt, & Sah, 2015; Szot, 2016;), and the neurotransmitter, norepinephrine, has been seen to be lacking in depressed suicide victims (Khan *et al.*, 2016; Klimek *et al.*, 1997; Ramirez, 2016). Norepinephrine levels have also been linked to studies on two types of creative people (Zabelina *et al.*, 2015), and in them both, a reduction in their aural sensory processing, known as sensory gating is tied to the neurotransmitter, while creative achievers have shown "leaky" sensory gating because they simultaneously focus on a large range of stimuli (Zabelina *et al.*, 2015). They have low levels of norepinephrine which increase the size and distribution of the brain's concept representations. Their ability to modulate the frontal lobe-locus coeruleus system and reduce norepinephrine levels leads to the discovery of novel orderly relationships, or creative innovation (Heilman, Nadeau, & Beversdorf, 2003). The other type of creative people, divergent thinkers, on the other hand, reduce sensory gating, which is also a marker of psychosis, including schizophrenia (Zabelina *et al.*, 2015). They have high levels of norepinephrine that restricts their concept representations (Heilman, Nadeau & Beversdorf, 2003), and therefore their sensory processing narrows to focus tightly on the task at hand. Here, thought influences ideas and modified aspects of the sensory cortex that feeds language.

8.2.3 Preparing the text

A collective corpus of 180,000 words contains a 104,000-word sample from 26 Iris Murdoch novels and a 76,000-word sample from 19 P.D. James novels (see Table 50 and Table 51). The 4,000-word novel sample is from the first 3,000 words and the last 1,000 words. Generally, in a novel, this is where characters, rich in setting and plot, are introduced and at the end, a conclusion of the general novel 'problem' has been resolved and summarised. We process the files as per Section 3.1. We also aggregate the data from the 45 POS types into 12 more general POS types, representing higher classes of Nouns, Verbs, Adverbs, Adjectives, Modal Verbs, Conjunctives, Prepositions, Determiners, Pronouns, Existential There, Articles, and Other categories, comprising Cardinal Numbers, Interjections, and Foreign words.

8.3 Analysis

In this section, we begin by testing for markers within the Richness (R) of language that can highlight Alzheimer's disease (AD), and we support this with Mann-Whitney U-Testing. We use Parts of Speech (POS) analysis to group the data by Content and Function Words and use ratios to support the presence of AD in the data further, and we back up this claim with Mann-Whitney U-Testing. We test for markers within the Sensory (S) aspects of language to see if the variables can identify additional markers for AD, and support these results with Mann-Whitney U-Testing and Principal Component Analysis.

8.3.1 Testing for Alzheimer's disease markers in Richness

Alzheimer's disease is apparent through lexical repetition, marked by smaller, higher frequency vocabulary 10-12 years before the disease is diagnosed (Garrard *et al.*, 2005; Rajan *et al.*, 2015). Iris Murdoch was diagnosed with Alzheimer's disease in 1997 (aged 78), and she died two years later in 1999, just four months before her eightieth birthday, and a post-mortem confirmed Alzheimer's disease (Garrard *et al.*, 2005). P.D. James was not diagnosed with Alzheimer's disease or dementia and died in 2014 (aged 94).

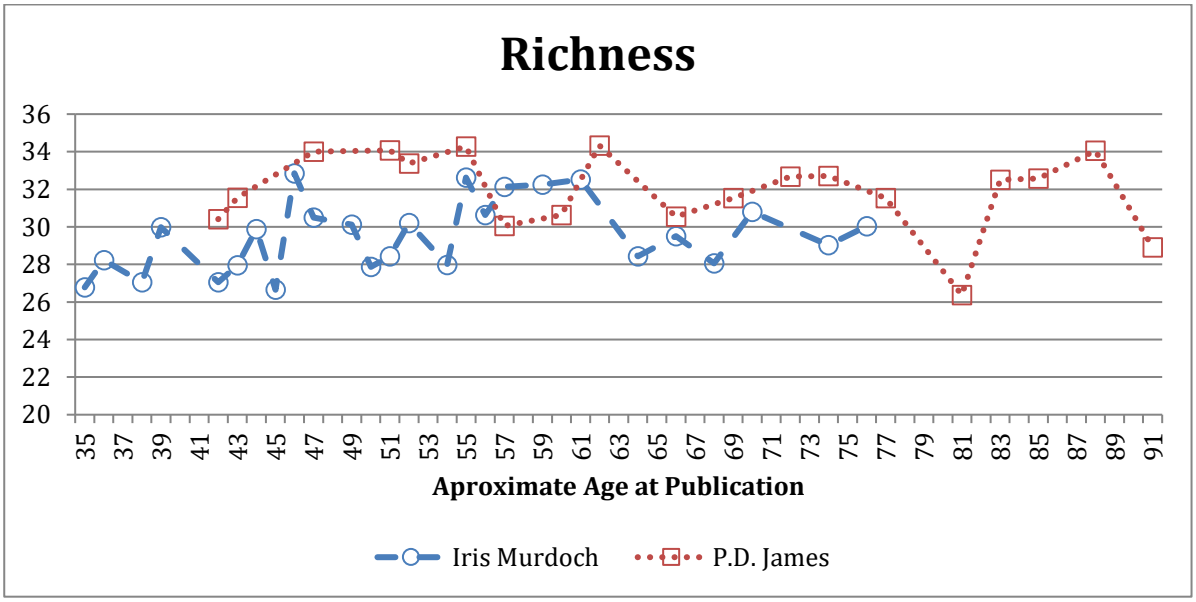
As can be seen in Richness (Figure 18), there appears to be a trough in Iris Murdoch's Richness scores marked by a period covering 9 years between age 46 to 55 (B9 - B16). Her Richness scores seem to climb through her writing career from the age of 35 to 76 (B1 - B26) so that her last book is 12.13% richer than her first. If we consider the point

where she was 61 years of age (B20), her word richness is higher again, and is an approximate 21.47% increase in unique word use, before it falls between ages 61 and 76 (B20 - B26).

If not for an anomaly at the age of 81 (B15), much of P.D. James scores are more consistent overall. As can be seen, the results are relatively flat. Similarly to Iris Murdoch, P.D. James has a trough marked by a period covering 7 years between age 55 to 62 (B6 – B9). In P.D. James' case, there appears to be a decline in her Richness scores during her writing career, marked by B15 and B19, and her last book is approximately 5.2% less rich than her first. If we ignore her final book, however, (B19), then her Richness score is overall higher, and it grows between the ages of 42 and 88 years of age by approximately 12%.

The total mean of Iris Murdoch's writing is 29.5, while P.D. James' is slightly higher at 31.9, suggesting there is less lexical repletion than Iris Murdoch's. Separating the last 12 years of works highlights that this period in Iris Murdoch's case is slightly lower (29.48 versus 29.52), while P.D. James' is slightly higher (32 versus 31.86) and these do not appear to be significantly different.

Figure 18: Richness by Age at Publication



To test this, we separate each author's writing into two groups, so that we can compare any changes in their last 12 years with earlier writing. We conduct a Mann-Whitney U-Test on the Richness scores. In this hypothesis testing for differences in a person's writing style, we use the Mann-Whitney U-Test because it is a non-parametric independent groups test. In this case, the total sample size for Iris Murdoch is 26, with

the group sizes of 21 and 5, and for P.D. James, the total sample size is 19 with the group sizes being 15 and 4. This test is ideal for unequal group sizes that are small, have dissimilar variances, and a distribution that is not normal (Burns & Burns, 2012).

We find significant differences in the lexical repetition of Iris Murdoch's works (see Table 15 for the Mann-Whitney U test mean ranks of the two groups and Table 16 for the test statistic results) during the writing period 1984 – 1996, 12 years before her diagnosis when compared to the earlier period of writing 1954 - 1983 (U=12.5, p = .009). There are no significant differences in the lexical repetition of P.D. James works (see Table 17 for the Mann-Whitney U test mean ranks of the two groups and Table 18 for the test statistic results) during the writing period 1962 – 2001, 12-years before her death when compared to the period 2002 - 2011 (U=29.0, p = .920).

Table 15: Iris Murdoch Mann-Whitney U-test 12-year ranks

Ranks

	AD	N	Mean Rank	Sum of Ranks
Rank of RICHNESS by AD	1	21	15.40	323.50
	2	5	5.50	27.50
	Total	26		

Table 16: Iris Murdoch Mann-Whitney U-test 12-year statistics

Test Statistics[b]

	Rank of RICHNESS by AD
Mann-Whitney U	12.500
Wilcoxon W	27.500
Z	-2.605
Asymp. Sig. (2-tailed)	.009
Exact Sig. [2*(1-tailed Sig.)]	.006[a]

a. Not corrected for ties.

b. Grouping Variable: AD

Table 17: P.D. James Mann-Whitney U-test 12-year ranks

Ranks

	AD	N	Mean Rank	Sum of Ranks
Rank of Richness	1	15	10.07	151.00
	2	4	9.75	39.00
	Total	19		

Table 18: P.D. James Mann-Whitney U-test 12-year statistics

Test Statistics[b]

	Rank of Richness
Mann-Whitney U	29.000
Wilcoxon W	39.000
Z	-.100
Asymp. Sig. (2-tailed)	.920
Exact Sig. [2*(1-tailed Sig.)]	.961[a]

a. Not corrected for ties.

b. Grouping Variable: AD

8.3.2 Content and Function Word Analysis

It is generally understood that while the ratios of different word types is relatively uniform across age, sex, and level of education in normal speakers, that there is a lower use of Function Words over Content Words in people with dementia and different aphasia types that impact speech and language (Bird *et al.*, 2000). This is because sentences are less complex. Function Words contain little meaning and tend to hold sentence structure together. They are word types such as pronouns, articles, prepositions, and conjunctions. Content Words, on the other hand, tend to describe the message of a sentence through verbs, nouns, adverbs, and adjectives.

Given these differences in Function and Content Words (Bird *et al.*, 2000), we would expect that as a person develops dementia, there would be an increase in the use of Content Words, and in their Content to Function Word ratios. We test the use of Content Words by aggregating the 12 tagged Parts of Speech groups into Content Words and Function words (see Table 19 for Iris Murdoch and Table 20 for P.D. James).

We separate the last 12 years of works to compare the later writing to the earlier period. In Iris Murdoch, there is a mean increase in Content Words use of 3.16%, or 75.4 (2380.0 – 2456.3), and a mean decrease in Function Words use of 4.65%, or 75.4 (1619.1-1543.67) in the later 12 Years period. In contrast, in P.D. James', there is a mean increase in Content Words use of 1.77%, or 42.28 (2344.46-2386.75), and a mean decrease in Function Words use of 2.55%, or 42.29 (1655.53-1613.25) in the later 12 Years period (Table 21). Iris Murdoch's use of Content Words was approximately 44%, or 33.11 (42.29-75.4) larger in the period 12 years before her diagnosis of Alzheimer's disease.

Table 19: Iris Murdoch Aggregated Content and Function Word Ratios

WORK	Content Words	Function Words	Ratio
B1	2284	1716	0.751
B2	2307	1693	0.734
B3	2424	1576	0.650
B4	2284	1716	0.751
B5	2281	1719	0.753
B6	2439	1561	0.640
B7	2368	1632	0.689
B8	2227	1773	0.796
B9	2321	1679	0.723
B10	2407	1593	0.662
B11	2487	1513	0.608
B12	2404	1596	0.664
B13	2451	1549	0.632
B14	2516	1484	0.590
B15	2320	1680	0.724
B16	2478	1522	0.614
B17	2341	1659	0.709
B18	2427	1573	0.648
B19	2391	1609	0.673
B20	2461	1539	0.625
B21	2343	1657	0.707
B22	2432	1568	0.645
B23	2465	1535	0.623
B24	2517	1483	0.589
B25	2500	1500	0.6
B26	2481	1519	0.612

Table 20: P.D. James Aggregated Content and Function Word Ratios

WORK	CW	FW	Ratio
B1	2400	1600	0.667
B2	2445	1555	0.636
B3	2317	1683	0.726
B4	2425	1575	0.649
B5	2373	1627	0.686
B6	2375	1625	0.684
B7	2267	1733	0.764
B8	2307	1693	0.734
B9	2382	1618	0.679
B10	2279	1721	0.755
B11	2308	1692	0.733
B12	2340	1660	0.709
B13	2328	1672	0.718
B14	2369	1631	0.688
B15	2252	1748	0.776
B16	2371	1629	0.687
B17	2404	1596	0.664
B18	2358	1642	0.696
B19	2414	1586	0.657

Table 21: Function to Content Word Ratios

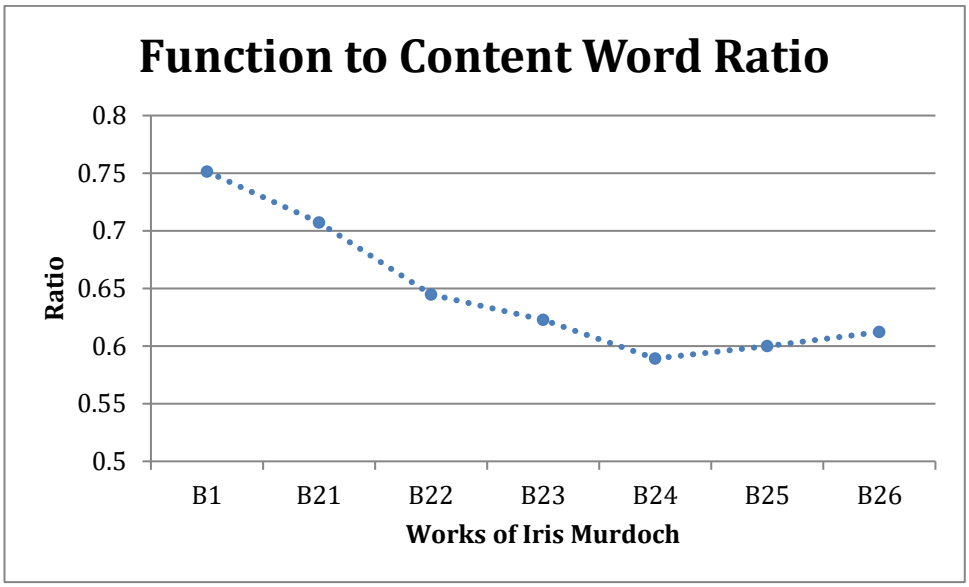
WORK	Content Words	Function Words	Ratio
P.D. James 4000 word sample			
B1	2400	1600	0.667
B16	2371	1629	0.687
B17	2404	1596	0.664
B18	2358	1642	0.696
B19	2414	1586	0.657
Iris Murdoch 4000 word sample			
B1	2284	1716	0.751
B21	2343	1657	0.707
B22	2432	1568	0.645
B23	2465	1535	0.623
B24	2517	1483	0.590
B25	2500	1500	0.6
B26	2481	1519	0.612

Having seen differences in Content Words, we test their Content to Function Word ratios for signs of dementia. Drawing on a technique from Garrard *et al.* (2005) and used in Le *et al.* (2011), we extend the approach by using a larger, more complete data set of the authors. We also look at ratios, which is a new approach (see Kavé & Goral,

2016). We compare the first work of Iris Murdoch to the period 12 years before her diagnosis of Alzheimer's disease. We plot the Function to Content Word ratio of Iris Murdoch and P.D. James (Table 21).

These results in Iris Murdoch's writing (Figure 19), reflect the findings of Garrard *et al.'s* (2005) observations that by her final book Jackson's Dilemma (B26), she was suffering from cognitive decline caused by Alzheimer's disease. Here, we extend the data from three of her novels (the first, middle and last) to seven (first and final six) and we find the Function Word to Content Word ratios are all lower for the six works 12 years before the diagnosis of AD. There is a steady decline in Iris Murdoch's work until the fourth work (B24) where the ratio is approximately level.

Figure 19: Iris Murdoch Content to Function Word Ratio Comparison of her first work to the six works 12 years prior to her diagnosis with AD. All variables are lower than her early work.



As we can see in P.D. James' writing (Figure 20), the Function Word to Content Word ratios are different from Iris Murdoch's. Here they appear as a sawtooth pattern, and while two are much higher (B16 and B18), the other two are lower (B17 and B19). There is only around 5.7% variation in P.D James' ratios (0.65 – 0.69), which suggests that there is neither a steady incline or decline in the 12 years before P.D. James' death. Iris Murdoch's variation is approximately 22.6% (0.58 – 0.75) which is a much higher level of variation. In Table 21 it is clear that Iris Murdoch's use of Content words increases, which is an indicator for AD, and this is seen in her use of Modal Verbs and Verbs generally.

These results are supported by the Mann-Whitney U-Testing, which show that there are significant differences in the use of Function Words and Content Words by Iris

Murdoch (U=24, p = 0.028) 12 years before her diagnosis of AD (see Table 22 and Table 23). As we would expect in a person without dementia or AD, there is no significant difference for P.D. James (U=17, p = 0.194) 12 years before her death (see Table 24 and Table 25).

Figure 20: P.D. James Content to Function Word Ratio Comparison of her first work to the four works 12 years prior to her death. The increasing sawtooth pattern is neither higher nor lower overall than her early work, and the results are very different to the Iris Murdoch results.

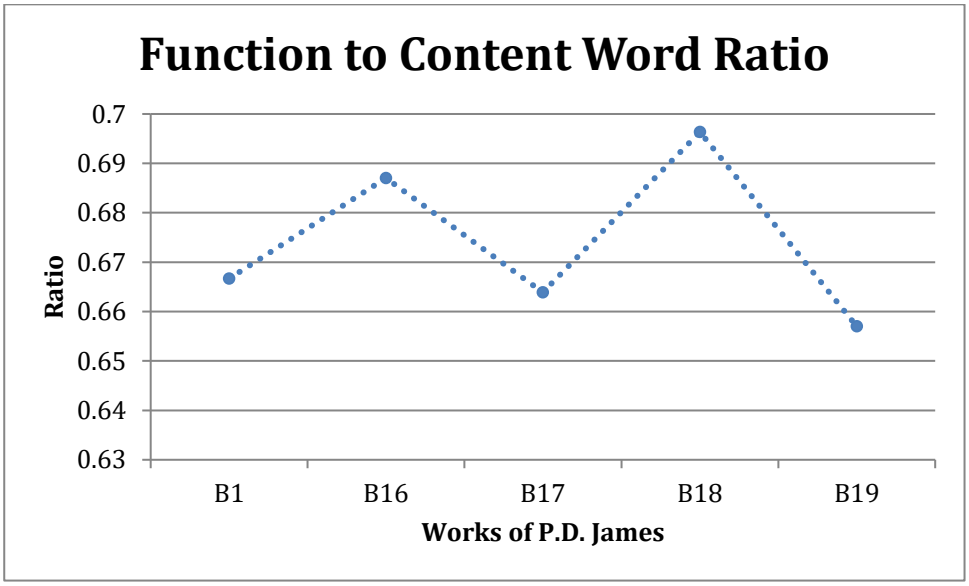


Table 22: Iris Murdoch Function to Content Word Ratio Mann-Whitney U-test Ranks

Ranks

	AD	N	Mean Rank	Sum of Ranks
Rank of Ratio	1	20	15.30	306.00
	2	6	7.50	45.00
	Total	26		

Table 23: Iris Murdoch Function to Content Word Ratio Mann-Whitney U-test Statistics

Test Statistics[b]

	Rank of Ratio
Mann-Whitney U	24.000
Wilcoxon W	45.000
Z	-2.191
Asymp. Sig. (2-tailed)	.028
Exact Sig. [2*(1-tailed Sig.)]	.028[a]

a. Not corrected for ties.

b. Grouping Variable: AD

Table 24: P.D. James Function to Content Word Ratio Mann-Whitney U-test Ranks

Ranks

	AD	N	Mean Rank	Sum of Ranks
Rank of Ratio	1	15	10.87	163.00
	2	4	6.75	27.00
	Total	19		

Table 25: P.D. James Function to Content Word Ratio Mann-Whitney U-test Statistics

Test Statistics[b]

	Rank of Ratio
Mann-Whitney U	17.000
Wilcoxon W	27.000
Z	-1.300
Asymp. Sig. (2-tailed)	.194
Exact Sig. [2*(1-tailed Sig.)]	.221[a]

a. Not corrected for ties.

b. Grouping Variable: AD

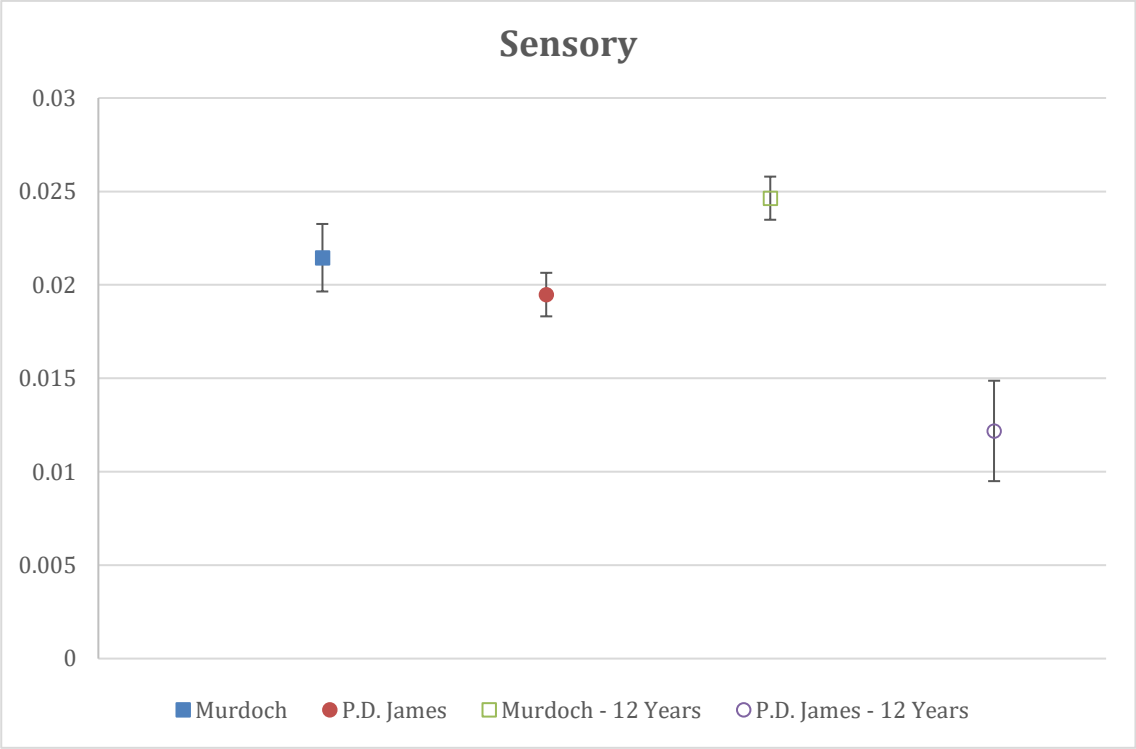
8.3.3 Testing for Sensory Alzheimer's disease markers

Many sensory words are processed by the brain as sight/feel and smell/taste word categories (Lynott & Connell, 2009), and we use a group of 387 sensory adjectives and allocate them a modality exclusivity score that reflects the brain's Representational System (van Dantzig *et al.,* 2011). These sensory words can be used to capture the sensory gating biomarker characteristics of a person (Fernandino *et al.,* 2015) to create a unique signature of their inner self.

Examining the differences between the earlier sensory variables to the final 12 years of Iris Murdoch's and P.D. James' writing (see Figure 21), we find that there is no overlap in the standard errors of their results. The mean of Iris Murdoch's writing is higher in the last 12 years, while P.D. James' Sensory variable 12 years before her death shows a lower sensory score. Overall, Iris Murdoch's Sensory mean is slightly higher than P.D. James (0.022 versus 0.018), but their earlier works are very similar. What is significant

are the opposing differences (Murdoch's higher and James' lower) that suggest a higher use of Adjectives in Murdoch's later writing.

Figure 21: Iris Murdoch and P.D. James' Sensory Mean with Standard Error bars highlighting Murdoch's higher use of sensory adjectives during the period 12 years prior to her diagnosis with AD. Note the error variance is smaller in the latter period and there are no overlaps in the standard error bars. In P.D. James' case, there is lower use of sensory adjectives during the period 12 years prior to her death. Note that these results are the reverse of the Iris Murdoch results in that the error variance is larger in the latter period with a lower mean. Again, there are no overlaps in the standard error bars.



These results are supported by the Mann-Whitney U-Testing, which show that there are significant differences in the use of Sensory Words by Iris Murdoch ($U=18$, $p = 0.011$) 12 years before her diagnosis of AD. This is also true of P.D. James ($U=7$, $p = 0.021$) 12 years before her death (refer detailed results Table 52 - Table 55 in Appendix A), but in P.D. James' case, her use of Sensory Words is lower, not higher.

Overall, these observations are further supported in the underlying five sensory variables for Iris Murdoch. Except for the Olfactory element, all of the other means of the sensory elements (V, A, H, & G) were higher in the period 12 years before her diagnosis of AD. However, there were overlaps in the Haptic and Olfactory Standard Errors. In contrast, all of P.D. James Visual, Auditory, Haptic, Olfactory, and Gustatory (VAHOG) elements were lower in the 12 years before her death, and there were no overlaps in the Standard Errors (see Figure 43 - Figure 52 in Appendix A).

Principal Component Analysis is initially conducted on RPAS using the five sensory Visual, Auditory, Haptic, Olfactory, and Gustatory (VAHOG) variables, so we have eight variables (RPA-VAHOG). Correlation analysis highlights 18 of the 56 possible correlations are in excess of 0.30, with many of them being quite strong. Both the KMO (.694) and Bartlett's tests (p=.001) produce criteria that support the application of PCA. Communalities varied between 0.851 and 0.535. Applying Kaiser's Rule and scree test, three factors are deemed important. Following rotation, factor one is loaded on the five sensory element variables and accounts for 31.9% of the variance. Factor two is loaded on the Richness and the five sensory element variables and accounts for 21.7% of the variance. Factor three is loaded on Personal Pronouns (Gender) and RA Power and accounted for 14.6% of the variance. Overall, the total variance explained by the three factors is 68.24%.

Table 26: Iterative PCA of the 45 words of Iris Murdoch and P.D. James, showing the effect of removing one of the sensory modalities

Element Removed	Total variance %	Bartlett's Test (p-value)	Impact on Variance (%)
Nil	68.24	0.001	-
Olfactory	46.38	< 0.001	21.86
Haptic	57.54	0.011	10.7
Visual	58.56	0.009	9.68
Auditory	61	< 0.001	7.24
Gustatory	65.3	< 0.001	2.94

Iterative PCA is conducted on the RPA variables, and in all cases, the total variance is reduced (varied between 58 - 60.9%) and Richness contributes the most to the variance (10.24%), highlighting its value as an indicator for differences. Iterative Principal Component Analysis is conducted on the VAHOG elements, and each one is removed and replaced one at a time to measure the effect it has on the total variance and to examine if it can be increased. In all cases, applying Kaiser's Rule and scree test, two factors were deemed important in every case except with Olfactory, which was limited to one factor. The total variance explained varied between 2.94 – 21.86%. As can be seen in Table 26, the order of contribution to the sensory variance is Olfactory, Haptic, Visual, Auditory, and Gustatory.

The works of Iris Murdoch and P.D. James are separated and iterative PCA conducted on each author's work to look at the large change in Olfactory impacts on total variance. We find that except for the Olfactory element, the results are relatively

similar across V, A, H, & G for both Iris Murdoch and P.D. James (see Table 27). In the case of the Olfactory element, we see a large, negative impact from its removal, highlighting the significant contribution, and therefore difference, which this element plays within the Iris Murdoch data. This large role played by olfaction is not seen in P.D. James work, where it's removal has a positive effect on the total variance explained within the data. Alzheimer's disease impacts normal olfactory function with suggestions that olfactory loss may be a biomarker for AD and cognitive decline (Wesson *et al.*, 2010; Woodward *et al.*, 2015).

Table 27: A comparison of the sensory contribution to PCA variance. Here it is clear that the impact on Iris Murdoch's Olfactory variable is significantly larger than any other result for either author.

Impact on Variance (%)		
Sensory Element	Iris Murdoch	P.D. James
Visual	-0.86	-1.13
Auditory	2.93	5.1
Haptic	-2.05	-1.3
Olfactory	-16.19	3.87
Gustatory	6.74	0.36

8.4 Discussion

Analysis of the Richness of each of the writer's novels using a Mann-Whitney U-Test highlighted significant differences in the lexical repetition of Iris Murdoch's works in the last 12 years of her writing (1984 – 1996). However, there were no significant differences in the works of P.D. James 12 years before her death. When using Parts of Speech analysis to group each work into Content and Function words., there was an increase in the use of Content Words in the later 12 Years of their writing. In Iris Murdoch's case, they were approximately 44% larger than P.D. James'. A decrease in Content to Function Word ratios as an indication of dementia was observed in the overall works of Iris Murdoch 12 years before the diagnosis of AD, with four of the six works declining before the ratios levelled. A Mann-Whitney U-Test supported the significant differences in the later period of her writing. No such decrease or significant difference was observed in the works of P.D James, and the rising and falling sawtooth variation showed a pattern of neither steady incline nor decline in the 12 years before her death.

While there are no prior documented links to the Sensory variable and dementia or AD, we test this and the five Sensory elements (VAHOG) for indications. We found that overall, Iris Murdoch's Sensory mean in the period 12 Years before her diagnosis of AD is higher, while P.D. James is lower than their earlier work, and these two groups are significantly different for both writers. Comparisons of the underlying five sensory variables for Iris Murdoch and P.D. James (see Table 28) highlight the means of the sensory elements (V, A, H, & G) were higher in the period 12 years before Iris Murdoch's diagnosis of AD. The Olfactory element was the exception. While some people have a poor sense of smell, in Iris Murdoch's case, her early olfactory scores were similar to P.D. James', and the difference lies in the comparison to her other sensory modalities, and how they were all different from P.D. James in the last 12 years of writing.

Table 28: Summary of Sensory Means of Iris Murdoch and P.D. James showing the overall higher sensory component in Iris Murdoch's results during the 12-year period prior to her diagnosis of AD, where P.D. James' results were all lower. Note that there was Standard Error overlap in Murdoch's Haptic and Olfactory scores and that the Olfactory value was equal, or only slightly higher than the earlier period.

Comparison of overlapping Stand Error bars between author's mean scores				
	Early Writing		Last 12 Years	
Variable	Iris Murdoch	P.D. James	Iris Murdoch	P.D. James
Sensory	No	No	Higher	Lower
Visual Sensory	No	No	Higher	Lower
Auditory Sensory	No	No	Higher	Lower
Haptic Sensory	Yes	No	Higher	Lower
Olfactory Sensory	Yes	No	Almost Equal	Lower
Gustatory Sensory	No	No	Higher	Lower

While olfactory dysfunction is seen in normally aging individuals, AD begins in the entorhinal cortex an area that affects the olfactory function and has the potential to be an early marker of neurodegenerative conditions such as AD, multiple sclerosis, schizophrenia and Parkinson's disease (Zou *et al.*, 2016). This change in olfaction is a physical one that is more significant than observed in normal aging. Looking at the observations from a linguistic perpective we see that all of P.D. James Visual, Auditory, Haptic, Olfactory, and Gustatory (VAHOG) elements were lower in the 12 years before her death. While there were overlaps in Iris Murdoch's Haptic and Olfactory Standard Errors, overall, the sensory elements except Olfactory were higher in the last 12 years prior to her diagnosis of AD. When conducting Principal Component Analysis and removing each of the five sensory elements of VAHOG one at a time we can see a significant observation in the Olfactory area of Iris Murdoch. The range of variation in

PCA analysis for P.D. James was -1.13 - 5.1 across all of the senses, while the range of variation in Iris Murdoch across all the senses (excluding Olfactory) was -2.05 - 6.75, which is quite similar. However, when the Olfactory element is removed the variation changes by -16.19, showing the high contribution that Olfactory words contribute to the variation in the data. Another words, the use of Olfactory words is significantly different in Iris Murdoch's writing (see Table 27).

Both Richness and POS analysis supported by Mann-Whitney U-Test highlight the evidence of dementia and AD in Iris Murdoch's writing in the latter 12 years before her diagnosis. These can be seen in Richness, a higher-level use of Content Words, and a lower and lower Content to Function Word ratios. We also find that a higher Sensory mean might suggest the presence of AD in the period 12 Years before Iris Murdoch's diagnosis, possibly because of her reliance on adjectives (and function words) rather than nouns. While this is supported by the analysis of the different means in the last 12 Years of writing and the earlier works, and the comparative differences in Principal Component Analysis variances, the exceptionally low Olfactory element compared to her other sensory modalities throughout her life and also observed in the last 12 years of her writing is quite different in Iris Murdoch's writing.

Murdoch's depression and apathy have been well documented through her prolific habit of writing about herself throughout her life (Dooley & Nerlich, 2014; Martin & Rowe, 2010; Murdoch, 2016; Wilson, 2004). Her decline into Alzheimer's disease has also been recorded by her husband (Bayley, 1998; 1999). We have stated earlier that there is a strong link between depression and apathy in dementia and particularly AD. It is known that while a depressed mood and apathy alter brain function in the prefrontal limbic network, that it overlaps regions dealing with olfaction, such that depression can reduce olfactory ability (Croy *et al.*, 2014).

A limitation to this longitudinal study is that it is the writing of only two authors and is not sufficient to suggest that AD, or indeed depression or apathy can be determined from the sensory writing of individuals. However, in this new approach to identifying the style of a person's writing using sensory adjectives, there were clear differences between both author's works in their last 12 years that warrant further study of other known authors who developed dementia.

8.5 Conclusion

In a study of two highly creative and prolific authors, we have been able to draw on a complete set of novels, more than that used by Garrard *et al.* (2005) and Le *et al.* (2011) to characterise a person's use of language through writing. In doing so, we have applied both known techniques to identify linguistic markers for Alzheimer's disease, depression, and apathy (through lexical repetition and function to content word ratios) and test a new technique based on sensory adjectives. Our results support the hypothesis that thought and language are impacted by depression and apathy and revealed in a person's writing style 12 years before a formal diagnosis of Alzheimer's disease present. Using Richness to measure lexical repetition (a form of the type-token ratio, Section 3.4.1.1) the writing of Iris Murdoch is statistically significantly lower in the last 12 years of her novel writing. This result is also reflected in Iris Murdoch's use of Function Words and Content Words). In contrast, a healthy P.D. James' writing during the same period showed no decline in lexical repetition and function to content word ratios and was not different to her earlier writing. There were clear differences in their use of sensory adjectives, with Iris Murdoch's use higher and P.D. James lower during their latter 12 years of writing, but in Iris Murdoch's case, her use of olfactory words, a biological sensory marker for Alzheimer's disease, depression, and apathy, was low. It is possible that olfactory sensory words in language could be used to help identify depression and apathy in people. We suggest that cognitive diseases such as dementia impact on thinking, as seen through depression and apathy and can influence language use.

8.6 Summary

In this study, the Parts of Speech analysis, Mann-Whitney U-Testing, and Principal Component Analysis identified both normal aging, and changes over time because of life events. The study demonstrated that the sample size used was sufficient to contain biomarkers for depression and Alzheimer's disease because the results were similar to other existing studies.

The most significant findings from this study were that when comparing a writer's earlier work to that 12 years prior to a formal diagnosis of AD, there are indications of lower Richness, increased content word use and a corresponding decrease in function words, another word, a falling Function to Content word ratio. With the exception of

Olfactory words, there was an increase in the use of sensory adjectives, suggesting that lower use of Olfactory words might be a marker for AD and cognitive decline.

RPAS Over Time and CSD Indicators

In this chapter, the third of the three studies looking at the changes across an individual's writing style over time is addressed, and it is the second of the two studies using the novels of Iris Murdoch and P.D. James. In the previous study (study five), the focus was on identifying both normal aging, and changes over time because of life events. In this study (study six), the aim is to use techniques to visualise the Critical Slowing Down phenomena and see if it is possible to identify any tipping points that might mimic a terrorist's mindset prior to an attack.

Using RPAS comparisons between Iris Murdoch and P.D. James, and the 1-lag autocorrelation and Fisher-Pearson coefficient of skewness techniques, analysis of the stylometric markers identified in the previous chapter are conducted to discover if a tipping point can be found in Iris Murdoch's writing and not in P.D. James'. To reinforce our earlier results and choice of sampling (study five), we conduct an independent comparative assessment of Iris Murdoch's writing style, prior to identified decline into Alzheimer's disease, using the Linguistic Inquiry and Word Count (LIWC) text analysis program. Using the techniques from the first study of the Elizabethan playwrights and poets (Chapter 4), Principal Component Analysis (PCA) and Stepwise Linear Discriminant Analysis (LDA) is conducted on this data to demonstrate that the techniques can separate the writing of contemporary authors, and not only 400-year-old text.

In this study, the third research question is addressed (Section 1.3) and hypothesis H_3 is tested (Section 1.4)

In this case, the findings are clear. Using 1-lag autocorrelation (AR1) and Fisher-Pearson coefficient of skewness (G1) techniques to measure the Critical Slowing Down (CSD) phenomena on the Sensory element, a tipping point was identified in Iris Murdoch. **Given these findings, we are able to reject the null hypothesis and say that the application of techniques to visualise the critical slowing down phenomena can identify changes in a person's moods, or shifts from one state to another, that might indicate a tipping point for self-radicalisation.**

This chapter is taken from a peer-reviewed paper: *The Stylometric Impacts of Aging and Life Events on Identity*. It was accepted for publication by the *Journal of Quantitative Linguistics.* (Accepted 12 November 2017).

9.1 Introduction

This paper is part of a wider study into the self-radicalisation problem using quantitative linguistic techniques to create a stylistic fingerprint of a person's personality – their personal signature – and reveal their 'identity' from their writing style. Here, we extend these quantitative linguistic techniques to determine if a person's 'identity' changes over time because of life events, such as trauma, depression, and disease, or if is it stable. Using the critical slowing down phenomena, a behaviour of complex dynamical systems (Dakos *et al.*, 2012; Drake & Griffen, 2010; Slater, 2013), on a person's writing to see if it is possible to identify changes in a person's moods, or shifts from one state to another, when a person is unable to cope with their environment, and whether this might indicate a tipping point for self-radicalisation. The issue of self-radicalisation is a problem that can be benefited through the application and analysis of linguistic techniques, in particular, stylometry.

While the rise of terrorism throughout the world is a key concern (Department of Defence, 2016). This threat is amplified when a terrorist act occurs within our neighbourhood by self-radicalised individuals with no known affiliation to terrorist organizations. No fully scientific theory has yet been developed to explain self-radicalisation or provide a predictive framework for current policy and operational needs (Reardon, 2015; Schiermeier, 2014), but terrorist organizations continue to exploit the use of social networking and other Internet sites by targeting Western people for recruitment, and radical material on social networking sites allows for self-radicalization, creating home-grown terrorists (Fuentes, 2016). The US government has no comprehensive strategy to combat the use of social media to radicalize potential terrorists, and law enforcement needs to better understand how to assess precursors of radicalization exhibited by potential terrorists (Layton, 2016).

With suicide terrorists, mental health problems, personal crises, coercion, fear of an approaching enemy, or hidden self-destructive urges play a major role in their actions (Lankford, 2014). Many of these illnesses are also prevalent in highly creative people. Creative writers, playwrights, and poets are believed to have a higher prevalence of

pathological personality traits, such as depressive disorders, bipolar affective psychosis, and alcoholism (Post, 1996).

We wonder if there is some linguistic markers within the writing of home-grown terrorists that might help support a predictive framework. It is believed that many mass murderer's lives are plagued with psychosis, paranoia, and depression, while lone actors typically suffer from mental illness and tend to be suicidal (Capellan, 2015). With suicide terrorists, mental health problems, personal crises, coercion, fear of an approaching enemy, or hidden self-destructive urges play a major role in their actions (Fuentes, 2016). However, despite a growing interest in the motivations and psychological profiles of suicide attackers, few empirical studies have examined their personal writings and recordings (Smith, 2016), and suicide terrorism is still a poorly understood phenomenon (Cohen, 2016). While suicide terrorists kill more people on average overall, non-suicide attacks can be just as lethal, but suicide operations are laden with symbolism and significance (Mroszcyk, 2016) The declared intention to die in order to kill others turns a suicide attacker into a powerful, highly dangerous, and utterly unpredictable weapon (Filote, Potrafke, & Ursprung, 2016).

Suicidality, the behavior related to contemplating, attempting, or completing suicide generally, appears higher in sexual minority youths than their heterosexual peers (Bostwick et al., 2014). However, the views on suicidality in attackers and whether attackers more generally are mentally ill is conflicting (Metzl & MacLeish, 2015). While many of the views held by the political science and international relations fields are that suicide terrorists are not suicidal, due to relatively few formal systematic studies of suicidality in suicide terrorists, there is emerging evidence to suggest that suicidality may play a role in a significant number of cases (Sheehan, 2014). Many agree there is a notion of Durkheim's (1897) 'altruistic suicide' where attackers motives are driven by a higher selfless act for the greater good (Pape, 2008; Nilsson, 2017).

In a study of 119 lone-actor terrorists and a matched sample of group-based terrorists, the odds of a lone-actor terrorist having a mental illness is 13.49 times higher, and those with a mental illness were more likely to have a proximate upcoming life change and experienced proximate and chronic stress (Corner & Gill, 2015; Meloy & Gill, 2016). The results identify behaviours and traits that security agencies can utilize to monitor and prevent lone-actor terrorism events (Corner & Gill, 2015). We believe that by drawing on data from creative writers, including ones with known depression and cognitive decline throughout their life, that it might be possible to mimic some of the

inner turmoil faced by suicide terrorist's face (Hoffman, 2017; Bhui, 2014; Speckhard & Akhmedova, 2006) using stylometric analysis techniques. We draw on the research conducted in Chapter 8.

We use **RPAS** to create individual stylistic signatures of the 45 Iris Murdoch and P.D. James novels. From simple comparative graphs, and a modification on the coefficient of skewness and 1-lag autocorrelation equations used in the Critical Slowing Down (CSD) dynamical property, we can observe the impacts of aging events on identity.

9.2 Methodology

We achieve a 180,000-word sample by taking 4,000 words from each of Iris Murdoch's 26 fiction novels (see Table 50) and 19 of P.D. James' fiction novels (see Table 51). We pre-process the data (Section 3.1). These 104,000 words from Murdoch's and 76,000 words from James' writing were taken from the first 3,000 words from the beginning of each novel and the last 1,000 words once punctuation is removed.

We had hoped this approach would capture the narrative structure of the story at a point in the introduction where characters, rich in setting and plot, are introduced and at the end, a conclusion of the general novel 'problem' has been resolved and summarized. This would reduce our data variation to more readily help 'see' any subtle changes in writing over time. Gee and Grosjean (1984) suggest sentences have a spatial aspect where they are related to sentences proceeding and following them, and that have a temporal element, where each one flows after the other in time, and that this narrative grouping is intuitively logical. However, Lehnert's (1981) model of story structure, suggests narrative is nothing more than simple plot units combined and connected to make up complex plot configurations. Books are not generally written from beginning to end without different areas being revisited and words added, so the temporal and spatial elements can be treated as one within the confines of a book's overall creation.

To test if we have captured a part of the narrative structure that is different from the rest of a novel's structure, we use the Linguistic Inquiry and Word Count (LIWC) text analysis program (Pennebaker *et al.*, 2015) to segment Iris Murdoch's *The Sandcastle* into 38 3,000 word chunks. Using eight of the commonly occurring parts of speech (adjectives, adverbs, articles, auxiliary verbs, conjunctives, prepositions, pronouns, and verbs) we compare the first 3,000-word segment (the one we used) to the remaining 37 segments. Mann-Whitney tests show that there is no significant differences in our

sample compared to any of the other 3,000 word samples in the novel (adjectives (U=10, p=.438), adverbs (U=14, p=.681), articles (U=1, p=.110), auxiliary verbs (U=0, p=.091), conjunctions (U=1, p=.110), prepositions (U=0, p=.092), pronouns (U=5, p=.218), and verbs (U=2, p=.132)). Perhaps as Brewer (1984) says, there is structure to story, but its scope is not set in any particular size. Therefore, we cannot be certain if we have captured a particular part of the narrative structure, but we do know that this study draws on previously reported data that is known to contain markers for Alzheimer's disease and cognitive decline (see Chapter 8).

9.2.1 Existing Alzheimer's disease (AD) markers

An in-depth discussion on AD markers can be found in Section 8.2.1, and this current chapter draws on much of the work in Chapter 8. Our data is representative of the full novels as demonstrated by the similarities of our findings using the data when comparing it to the findings of Le *et al.* (2011) and Garrard *et al.* (2005). Further, we use Parts of Speech Tag analysis of Content and Function word ratios to highlight the cognitive decline in Iris Murdoch's writing. Using well-documented Parts of Speech (POS) techniques from a number of these recent articles we show that this data sample is robust enough to highlight markers for dementia and in particular, Alzheimer's disease, 10-12 years prior to diagnosis (refer to Chapter 8).

Here, we apply the RPAS multivariate technique to determine if we can further separate the effects of natural aging from Alzheimer's disease, which is something that van Velzen, Nanetti, and de Deyn (2014) have raised as a valid concern missed within these prior Alzheimer studies. We use the 19 novels of P.D. James as a control author because unlike Iris Murdoch, there were no indications she had any symptoms of Alzheimer's disease prior to her death.

9.2.2 Critical slowing down (CSD) visualisation

The approach taken with critical slowing down, including the two equations used, can be found in the Methods Section 3.4.2.

9.3 Analysis

In this section, we begin by making observations of the four RPAS elements and compare the works of Iris Murdoch against P.D. James using their approximate ages based on the novel publication dates. In reality, many of P.D. James novels were written over a period of three years prior to publication, while Iris Murdoch was

generally a quicker writer, producing novels around one to two years prior to publication. We then focus on the results of the Sensory element and conduct Critical Slowing Down (CSD) visualisation using the Skewness and 1-lag autocorrelation techniques.

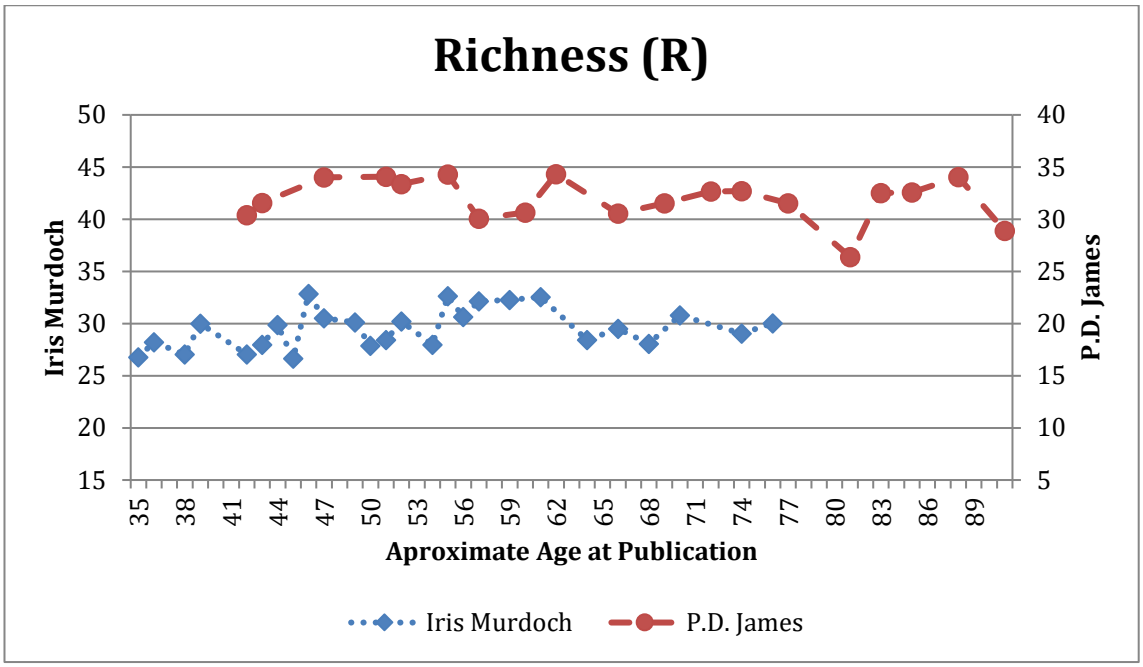
9.3.1 RPAS visualisation

Richness (R). In Figure 22, there appears to be a trough in Iris Murdoch's Richness scores marked by a period covering 9 years between age 46 to 55 (B9 - B16). Her Richness scores seem to climb through her writing career from the age of 35 to 76 (B1 - B26) so that her last book is 12.13% richer than her first. If we consider the point where she was 61 years of age (B20), her word richness is higher again, and is an approximate 21.47% increase in unique word use, before it falls between ages 61 and 76 (B20 - B26).

If not for an anomaly at the age of 81 (B15), much of P.D. James scores are more consistent and flat overall. Similarly, to Iris Murdoch, P.D. James has a trough marked by a period covering 7 years between age 55 to 62 (B6 – B9). In P.D. James' case, there appears to be a decline in her Richness scores during her writing career, marked by novels B15 and B19, and her last book is approximately 5.2% less rich than her first. If we ignore her final book, however (B19), then her Richness score is overall higher, and it grows between the ages of 42 and 88 years of age by approximately 12%. Overall, P.D. James has a higher Richness score than Iris Murdoch.

The unusual dip toward the end in the Iris Murdoch data, before falling into decline prior to her being diagnosed with Alzheimer's disease, is highlighted as a key AD marker in Le *et al*. (2011).

Figure 22: Richness (R) by Age at Publication for Iris Murdoch and P.D. James showing James' higher scores



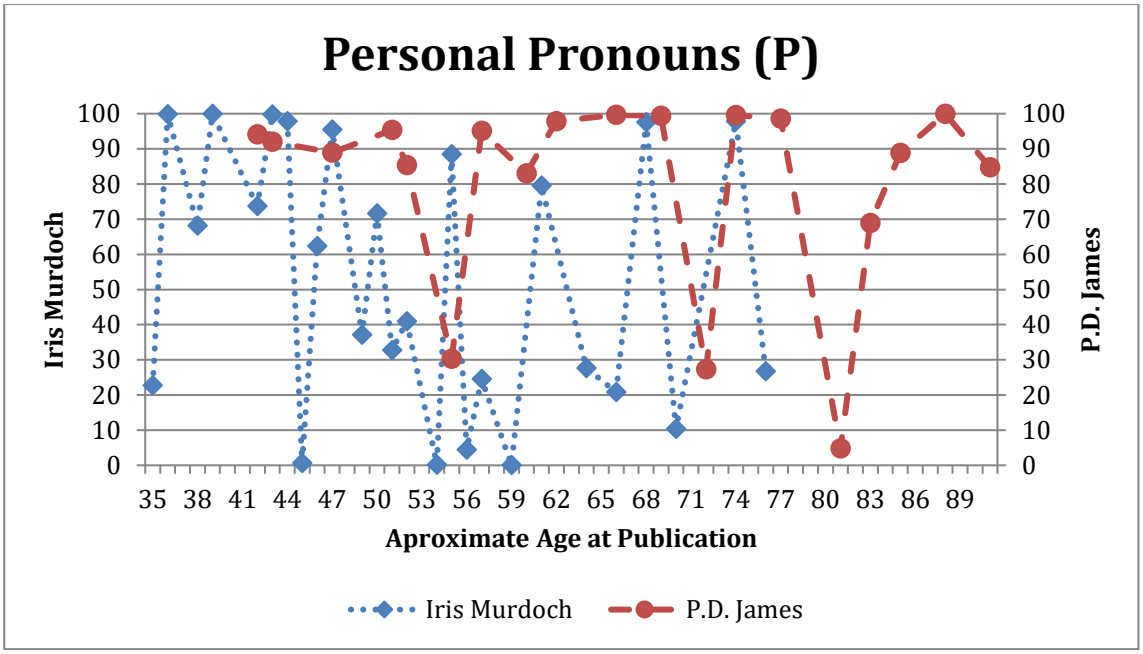
Personal Pronouns - Gender (P). In Figure 23, there is a decline in Iris Murdoch's scores marked by a period covering 7 years between the age 47 to 54 (B10 –B15). With the exception of B16, there is what appears to be a trough covering 14 years between the age 47 to 61 (B10 - B19). After that point, there is a continuous switching high and low during her writing career from age 66 (B22 - B26). There is also a major transition from age 44 to 45 (B7 - B8). Overall, there is a declining trend in the use of Personal Pronouns over Iris Murdoch's 26 novels, and the low points at B8, B15, and B19. Between the low points B15 and B19, there is the large transit at B16 and during this time Iris Murdoch was troubled by her mother's illness. The point prior to the earlier low point at B8 was preceded by another difficult time in her life when she had to leave St Anne's College. All three novels (B8, B15, and B19) are told from a male narrator's perspective and highlight the use of female personal pronouns 'my' and 'her' over the male use of 'its'.

P.D. James in comparison has fewer transitions, but there are three consecutively larger dips at age 55, 72, and 81 (B6, B12, and B15). Otherwise, most of the variables are very similar and quite high in value.

Unlike Iris Murdoch, there is no a constant steady trend in the use of Personal Pronouns over P.D. James' 19 novels. There is no publically available data to suggest any low points B6, B12, and B14, outside of initial novels prior to B6. Like Iris Murdoch, all three novels are told from a male narrator's perspective and highlight the

use of female personal pronouns 'my' and 'her' over the male use of 'its'. From age 77 during the rest of P.D. James' writing career, the gender pronouns tended to mimic the Richness results above. Overall, P.D. James has a higher Gender Pronoun score at 80.72%, while Iris Murdoch's is lower at 53.12%.

Figure 23: Personal Pronouns – Gender (P) by Age at Publication for Iris Murdoch and P.D. James showing James' higher and more consistent scores.

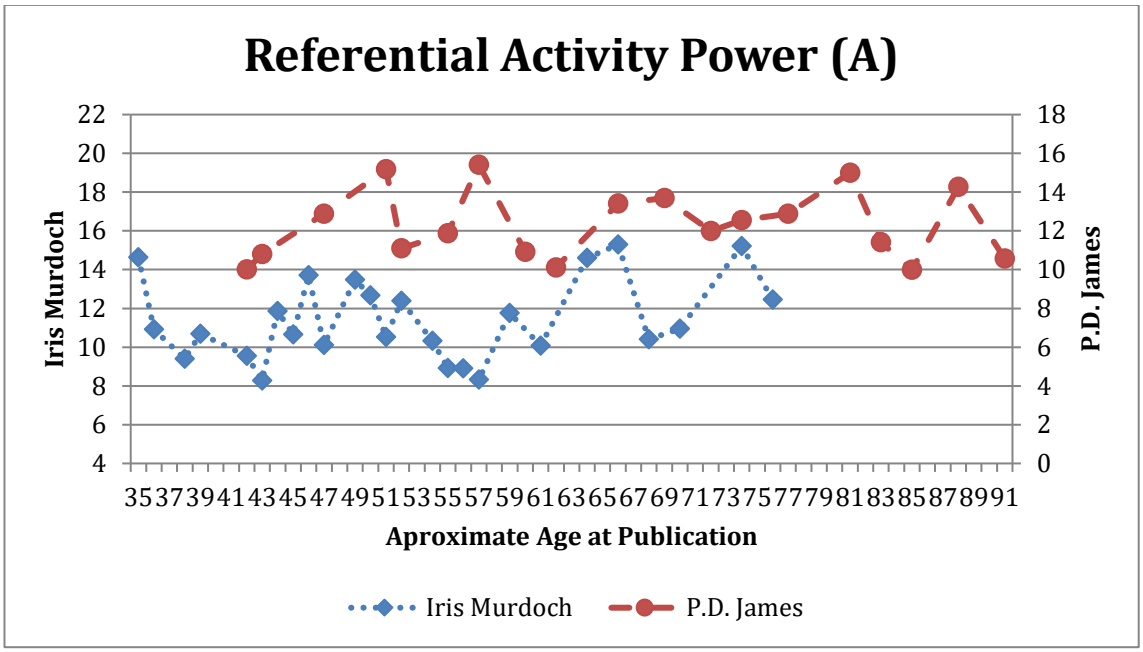


RA Power (A). In Figure 24, there are two low points in Iris Murdoch's RA Power scores at age 43 and 57 (B6 and B18). There seems to be a decline from age 39 to 43 (B4 - B6), and possibly overall from age 35 to 43 (B1 - B6) if the low peak at age 39 (B4) is discounted. There also seems to be a decline from age 52 to 57 (B14 - B18), and possibly overall from age 49 to 57 (B11 - B18) if the low peak at age 52 (B14) is discounted.

At 42 years of age, P.D. James commences at a low point (B1) and shows a more consistent mapping. There are two other low points at age 62 and 85 years of age (B9 and B17). At age 85 (B17), RA Power rises before dropping lower again. While Iris Murdoch's work tends to transition to low points over many iterations of her work, P.D. James' seems to have more iterations of climbing to a high point. There appear more transitions and variation in Iris Murdoch's work over her writing career than P.D. James. Overall, P.D. James has a slightly higher Referential Activity Power score, sitting above 10, while Iris Murdoch's falls below at B3, B6, and B16-18. When the two author's scores are compared, Murdoch's levels at scores below ten correspond to

documented low points in her life where is struggling with depression, anxiety and, personal trauma.

Figure 24: Referential Activity Power (A) by Age at Publication for Iris Murdoch and P.D. James showing James' higher and more consistent scores between 10-16, and Murdoch's three groups falling below ten. B3 at age 38, B6 at age 43, and B16-B18 at ages 55-57 mark documented low points in her life.



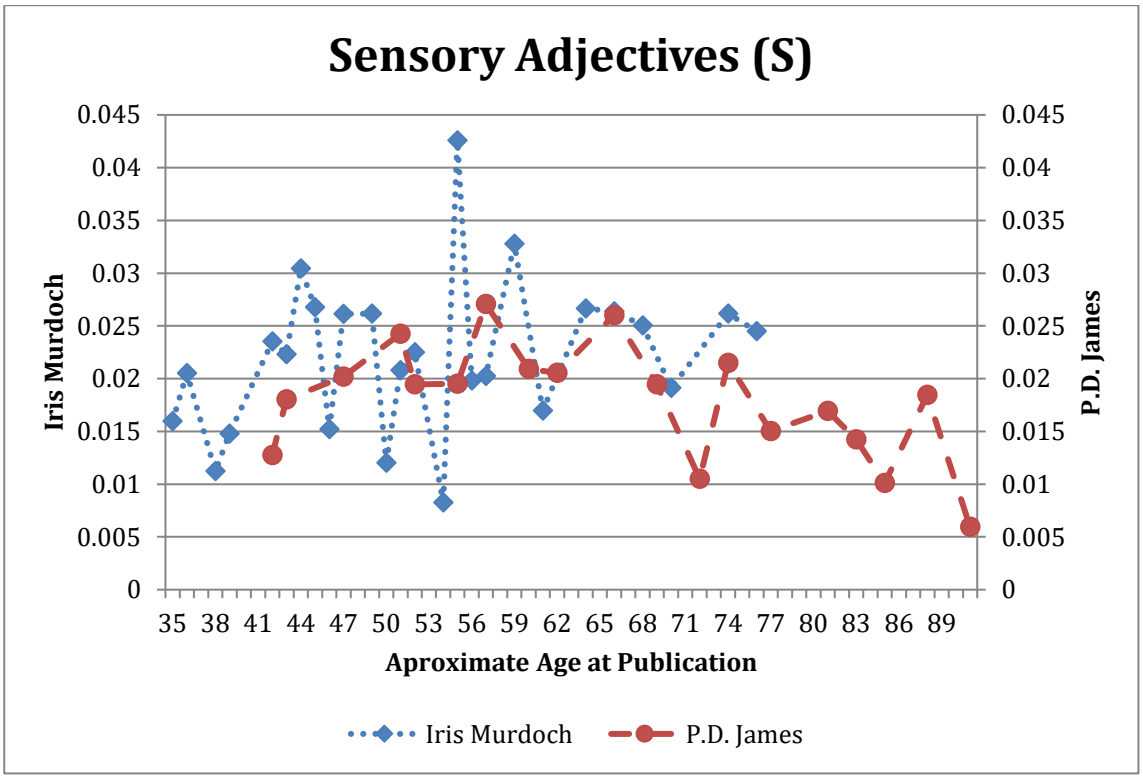
Sensory Adjectives (S). In Figure 25, there are two major sensory transitions in Iris Murdoch's work between the age of 54 to 56 (B15 - B16 and B16 - B17). These transitions are much larger than P.D. James' ones (B15 - B16 is 0.0343 compared with P.D. James' combined score of 0.0155 for B10-B12 at age 66 to 72 years). Iris Murdoch's Sensory indicators seem to increase by approximately 52.31% during her 41-year writing career.

At 42 years of age, P.D. James commences at a low point (B1). There is a low point at age 72 (B12) and another lower again at the age of 91 (B19). Overall, the Sensory indicators seem to fall during her 50-year writing career by approximately 114.65%, and this is mainly due to the final work at age 91(B19). If B19 is ignored, then the Sensory indicators appear to increase by approximately 44.65%. There is an overall decline from ages 74 to 91 years of age (B13 - B19).

Between the time when they commenced publishing novels up to a point at age 74, both Iris Murdoch's and P.D. James' sensory elements have similar characteristics of highs and lows (ignoring the large transition at age 54 to 56,

B15- B16- B17), and with an overall increase. There is a decline seen from the age of 74 to 91 in the work by P.D. James, and of note, there is a similar pattern of decline seen in the work of Iris Murdoch's from age 64 until she ceased writing at age 76, although it is not as pronounced as P.D. James (~8.83% versus ~262%) and is of shorter duration (12 years versus 17 years). Overall, Iris Murdoch's work has a higher Sensory score, even when B16 at age 55 is removed.

Figure 25: Sensory Adjectives (S) by Age at Publication for Iris Murdoch and P.D. James showing James' overall consistent decline and Murdoch's frequent transitions including B15-16 at age 54. The transition to B16 marks a major life event, a point of known difficulty in Murdoch's life where she was nursing her dying mother.



9.3.2 CSD visualisation

The sensory scores were processed using both an adjusted 1-lag autocorrelation technique (AR1) and an adjusted Skewness (G1) technique. A zero score was added to the end of the AR1 data, so this final variable can be ignored with little change in results. The results of both techniques have been shifted by the lowest negative value to display them in the positive quadrant.

Looking at the AR1 results first, we see two tipping points where the data climbs steadily over a number of accumulated samples and then falls (Figure 26). With the Iris Murdoch accumulated data, it climbs to a peak B1-B7 and then falls through B8-B15 (a

tipping point at B7 prior to the change). The remaining 11 accumulated samples are similar and appear as a flat line. What is interesting is that the flat line B3-B4 and B15-B26 mark times when Murdoch was also troubled. In comparison, P.D. James work falls through B1–12, but then climbs B15 and then falls through B16-B19 (a tipping point at B15 prior to the change). Unfortunately, P.D. James' has no documented life event that might explain this, but she was noted for keeping her personal life private.

Figure 26: 1-lag autocorrelation (AR1) of the accumulated 26 Iris Murdoch novels and 19 accumulated P.D. James novels. In Murdoch, we see a tipping point at point B7, where the rising autocorrelation falls. It rises with a decline through to point B15 where it flat lines (from age 54). This point at 15 and the transition at 3-4 are two other known times where Murdoch was troubled. In the corresponding P.D James' dataset, we see also see a rising trend between 12 and 15, with a tipping point at 15 (from age 81).

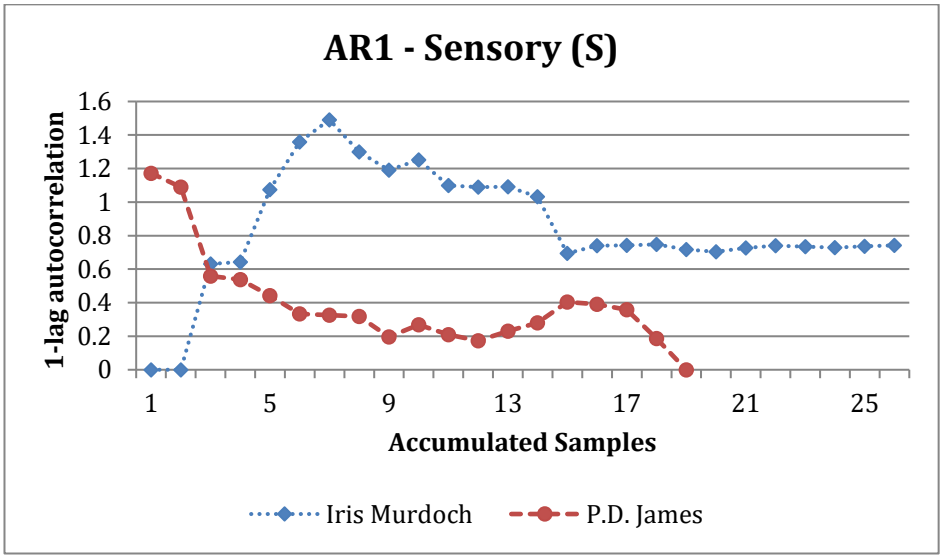
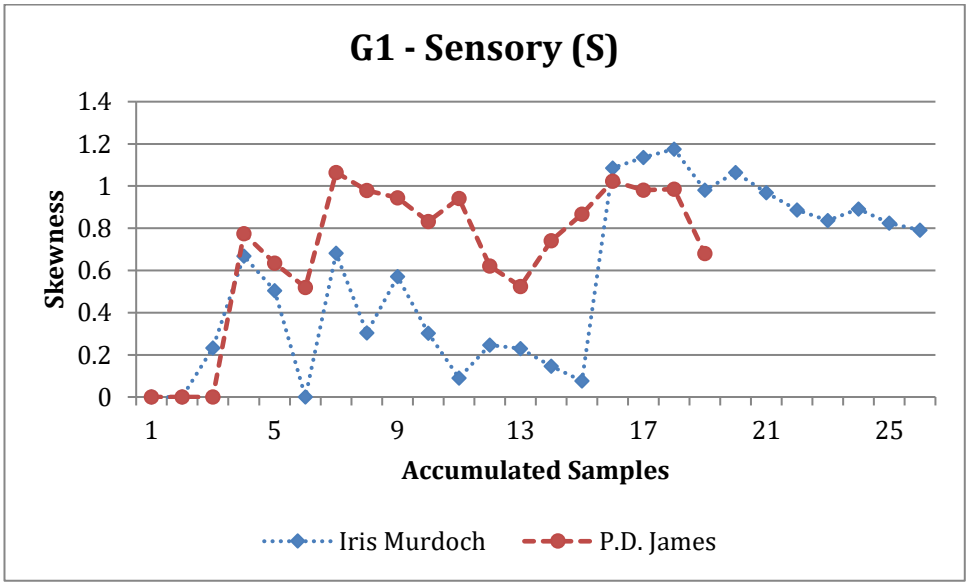


Figure 27: Skewness (G1) of the 26 Iris Murdoch and the 19 P.D. James works. In Murdoch, we see many transitions before a tipping point at point B18, while in James' works we see an overall rise and fall with a possible tipping point at B16.



Looking at the G1 results, we see two tipping points where the data climbs steadily over a number of accumulated samples and then falls (Figure 27). With the Iris Murdoch accumulated data, it climbs to a peak B15-B18 and then falls through B19-B26 (a tipping point at B18 prior to the change). There is a lot more variation in the data compared to the AR1 technique. In comparison, P.D. James work falls through B13–B16, but then falls through B17-B19 (a tipping point at B15 prior to the change) and close to the AR1 results.

9.4 Discussion

When looking at the results of the RPAS visualisation a number of points can be made. Overall, Iris Murdoch used fewer unique words and had more repetition than P.D. James which can be an indicator for dementia and Alzheimer's disease (Bird *et al.*, 2000; Garrard *et al.*, 2005; Lancashire & Hirst, 2009). P.D. James used more consistently male gendered pronouns in all her novels (~80%), and it didn't matter whether the main character was Adam Dalgliesh or Cordelia Gray (see B5, B9). It is possible that Iris Murdoch's Personal Pronoun score (~53%) reflected her well documented sexuality (Murdoch, 2016). P.D. James Sensory scores were lower overall, indicating that she drew less from her emotional experiences and used less imagery, while her writing concepts were more vague and abstract. P.D. James' work was also more consistent with less variance in her RPAS results. In comparison, Iris Murdoch works had more and larger transitions. Two clear differences were observed in the Referential Activity Power score and Sensory results.

Overall, P.D. James has a slightly higher Referential Activity Power score, sitting above 10, while Iris Murdoch's falls below at B3, B6, and B16-18. These might indicate negative life events. At the low point in 1957 (B3 aged 38), after the death of her father, Iris Murdoch admitted to being depressed and she abandoned her unpublished novel, *Jerusalem* when it was given a bad review by her lover and mentor, her depression fuelled by the varied reception of her earlier novels (Martin & Rowe, 2010; Murdoch, 2015). At the low point in 1962 (B6 aged 43), Murdoch was torn between her desire for a number of male and female lovers and the security of her married life, but was asked to leave her teaching post at St Anne's because of an inappropriate 'incident' with a fellow female staff member (Murdoch, 2016; Wilson, 2004). At the low point during 1974 - 1976 (B16-B18 aged 55-57 years), Iris Murdoch highlights that she has been dealing with a host of problems, her mother's mental health problems and diagnosis of dementia, requiring Iris's mother to stay with her while Iris took on a large burden of

care (Dooley & Nerlich, 2014; Murdoch, 2016; Wilson, 2004). We also know that P.D. James struggled with a schizophrenic husband during 1962-1963 during her first two novels (B1-B2 aged 42-43 years), and although for the most part he was locked away in an asylum, he did have regular home visits before he finally killed himself in the family home on their 24th wedding anniversary (Wilson, 2014). Her third book was written three years after his death, in 1967 (B3 aged 47 years), and tracks in the positive direction.

Looking at the Sensory elements and the Critical Slowing Down visualisation, we see that overall, Iris Murdoch's Sensory scores were higher, indicating that she drew more from her emotional experiences and imagery, and her concepts were less vague and less abstract than P.D. James. What is interesting is the major transition at B15-16 which is reinforced by the low Referential Activity power scores at the same period. It has also been stated that Alzheimer's disease (AD) can be seen in people's writing 10-12 years before the disease is diagnosed, and it has been suggested that in a few cases this might extend out to 18 years (Rajan *et al.*, 2015). Given that we had identified markers for Alzheimer's disease 10-12 years out prior to Iris Murdoch's diagnosis in the Richness scores, and found no markers in P.D. James in the same period, we suggest that the rising peak on the G1 chart is a tipping point at B17 (Figure 27) that rises from B15.

Further, given the AR1 flat line observation (Figure 26) at B15 in 1973 (aged 54), we wonder if this might be a linguistic marker for AD 20 years prior to her last novel when it had become clear her writing was impacted by AD, and given there were no corresponding indications in the results from P.D. James which suggest her steady decline observations in AR1 (Figure 26) could be from normal aging.

There are several differences in this approach compared with other earlier studies. Garrard *et al.*'s (2005) study highlighted significant lexical differences between Iris Murdoch's early work and her last novel. In their study, they used fully-parsed texts of both Murdoch's first novel, *Under the Net* (1954), the last novel, *Jackson's Dilemma* (1950), and the first 100 pages of her awarded novel, *The Sea, The Sea* (1978). Le *et al.*'s (2011) large-scale longitudinal study of Iris Murdoch and P.D. James used fully parsed texts and a larger number of measures to improve on the Garrard *et al.*'s (2005) study. Le *et al.* (2011) used 20 fully parsed Iris Murdoch novels and 15 of P.D. James' novels. In this study, we have taken a consistent 4,000 sample from all 26 of Iris Murdoch's fiction novels and all 19 of P.D. James' fiction novels. We can see the similar trough

identified in Murdoch's work by Le *et al.* (2011), and the lexical differences identified by Garrard *et al.* (2005).

We believe this study takes the above-mentioned results further. Using 45 works and multivariate techniques that include RPAS, and a variation on 1-lag autocorrelation and coefficient of skewness, we have tried to account for van Velzen *et al.*'s (2014) criticism of the impacts of natural aging on identity. We find a distinctive tipping point in the works of Iris Murdoch and P.D. James, and this could be a useful technique to apply to the CSD phenomena. However, what is interesting is the AR1 point 20 years earlier than Murdoch's formal diagnosis of AD (that compares to the extreme estimates in Rajan *et al.* (2015), and yet the results show what appears to be only signs of natural aging in the works of P.D. James.

While this stylometric approach shows merit, it is not sufficient to suggest that AD can be determined from the sensory writing of individuals using modified techniques within the CSD phenomena, but it warrants further study.

9.5 Conclusion

When combining **RPAS** with the modified 1-lag autocorrelation and coefficient of skewness, clear differentiation in the style of Iris Murdoch and P.D. James' writing can be seen. In addressing van Velzen *et al.*'s (2014) concerns about changes in a person's identity over time, these findings extend the finding of Rajan *et al.* (2015) by between 2-8 years. It is clear that a person's identity changes over time due to aging and life events, and we find that life events such as depression, anxiety, and Alzheimer's disease might be identified outside of natural aging through a tipping point phenomenon. We believe these techniques might be a useful self-help tool to aid in the signaling of depressive episodes, such as averting suicide, and the early identification of Alzheimer's disease, or for law enforcement personnel monitoring terrorists on watch lists.

9.6 Summary

In this chapter, the use techniques to visualise the Critical Slowing Down phenomena were effective at identifying a tipping point in the writing of Iris Murdoch. The most significant findings from this study were that the flat signal observed at Murdoch's 15th novel in the AR1 and G1 techniques could indicate the beginning of Alzheimer's disease at a point 20 years prior to a formal diagnosis. The falling Referential Activity

Power that corresponds to known difficult times in Iris Murdoch's life was also significant.

From this study, knowing that suicide attackers and lone wolf terrorists are more likely to have a mental illness, experience chronic stress, and have an upcoming life change, it is hoped this technique will be able to identify changes within them. Therefore, in the next chapter, the study of the final manifestos and notes of suicide attackers, we aim to attempt to improve the classification of a terrorist's theoretical writing, and separate their stylistic signature from both a normal person and somebody with depression who suffered from life changes see if we can separate their writing from normal blog posts.

Part Three: Terrorist Characterisation

In part three, the last in this research thesis, the focus is on improving the classification techniques identified in parts one and two.

In this part, the suicide notes and final manifestos of suicide attackers are compared to normal blog posts (Chapter 10) to see if RPAS is able to differentiate the writing. The works of Iris Murdoch are then added and compared, to see if a normal person with depression who is suffering from life events – something a terrorist and suicide attacker would also have – can be separated from a suicide attacker. In this study, the fourth and final research question is (hypothesis H_4): Can the final writings of suicide attackers be separated from 'normal' bloggers?

There is one paper that contributes to this in this third part (refer Section 1.6): *Identifying Suicide Attackers in Cyberspace*. More detail about it follows in the following chapter.

Suicide Attackers

In this chapter, the final study in this research thesis is addressed. The study draws on the suicide notes and final manifestos of suicide attackers, normal blog posts, and the works of Iris Murdoch, to see if RPAS is able to differentiate their writing and improve the classification techniques from the previous studies (Chapters 4 – 9). Iris Murdoch is chosen because of the known life events she suffered (Chapters 8-9), something a terrorist and suicide attacker may also have experienced.

Using the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker *et al.*, 2015), the analysis is conducted of the Suicide Attacker data to ensure it contains expected negative emotion and anger sentiment. Mann-Whitney U testing and Stepwise Multiple Regression Analysis are conducted. The findings are supported by 5-fold cross-validation techniques.

In this study, the fourth research question is addressed (Section 1.3) and we test hypothesis H_4 (Section 1.4).

In this case, the findings are clear. The nine elements of RPAS (including VAHOG) were used in step-wise multiple regression analysis, and four of the eight variables (RPAV) were statistically significant in predicting suicide attackers. **Given these findings, we are able to reject the null hypothesis and say that the final writings of suicide attackers can be separated from 'normal' bloggers.**

This chapter is taken from an accepted abstract and a paper presented at the Terrorism and Social Media Conference, in Swansea, Wales in June 2017, titled: *Identifying Suicide Attackers in Cyberspace.* It has been peer reviewed by the Studies in Conflict and Terrorism Journal.

10.1 Introduction

The rise of terrorism throughout the world is a key concern (Department of Defence, 2016). This concern is amplified when an act of terrorism occurs in the neighbourhood by self-radicalised individuals with no known affiliation to terrorist organizations. While no fully scientific theory has yet been developed to explain self-radicalisation, or

provide a predictive framework for current policy and operational needs (Reardon, 2015; Schiermeier, 2014), terrorist organizations continue to exploit the use of social networking and other Internet sites by targeting Western people for recruitment and creating home-grown terrorists (Fuentes, 2016).

It is believed that many mass murderer's lives are plagued with psychosis, paranoia, and depression, while lone wolves typically suffer from mental illness and tend to be suicidal (Capellan, 2015). With suicide terrorists, mental health problems, personal crises, coercion, fear of an approaching enemy, or hidden self-destructive urges play a major role in their actions (Lankford, 2014). However, despite a growing interest in the motivations and psychological profiles of suicide attackers, few empirical studies have examined their personal writings and recordings (Smith, 2016) and suicide terrorism is still a poorly understood phenomenon (Cohen, 2016). While suicide terrorists kill more people on average overall, non-suicide attacks can be just as lethal, but suicide operations are laden with symbolism and significance (Mroszczyk, 2016). The declared intention to die in order to kill others turns a suicide attacker into a powerful, highly dangerous, and utterly unpredictable weapon (Filote, Potrafke, & Ursprung, 2016).

Suicidality, the behavior related to contemplating, attempting, or completing suicide generally, but appears higher in sexual minority youths than their heterosexual peers (Bostwick *et al.*, 2014). While many of the views held by the political science and international relations fields are that suicide terrorists are not suicidal, and this is due to relatively few formal systematic studies of suicidality in suicide terrorists, but there is emerging evidence to suggest that suicidality may play a role in a significant number of cases (Sheehan, 2014). One study (Egnoto & Griffin, 2016) believe suicide terrorists are not suicidal, and in their study of three separate groups, the suicide notes, legacy e-mails, and social media text from people who commit multiple murders without a cooling-off period (known as spree killers), were compared to people who commit suicide without killing anyone else, and also to normal students. Their study highlighted that spree killers' negative emotions, and anger vocabulary use was significantly higher than normal students and those who commit suicide, but that the spree killers' use of personal pronoun and future tense vocabulary use was the same as normal students and higher in suicide victims.

This chapter is part of the wider thesis study into the self-radicalisation problem, and to date, it has been able to create a stylistic fingerprint of a person's personality – their personal signature – and revealed their 'identity' from their writing style (see Chapters

7 and 8). It has also determined that a person's 'identity' changes differently over time because of life events, such as trauma, depression, and disease, compared to someone who has not suffered the same way (see Chapter 6). By applying CSD techniques to visualise the tipping points (Section 3.4.2), it is possible to identify these changes in a person's moods, or shifts from one state to another, when a person is unable to cope with their environment, and these events might indicate a tipping point for self-radicalisation (see Chapter 9). Here, the earlier research is extended to determine if the final notes and manifestos of suicide attackers' writing are different from normal bloggers' online writing. There are linkages because lone wolf terrorists are more likely to have a mental illness or a proximate upcoming life change and experienced proximate and chronic stress (Corner & Gill, 2015). Many of them will also broadcast their intent through blogs (Cordy, 2017).

10.2 Methodology

The Global Terrorism Database or GTD (START, 2016) is used to identify 65 instances where a suicide attacker killed at least one other person before taking their own life. Using the 65 instances from the GTD, an open source search was then conducted for available open source data on the internet to gather as many suicide notes and final manifestos as possible. Only twenty-five English notes were found (average 1,704, 69-12,479 words), and many were handwriting images that needed transcribing into text files so they could be processed. Five web posts were added to an existing anonymous randomly collected data sample of 30 web posts and online articles (see Kernot, 2013).

The suicide attacker data comprised of suicide notes and final manifestos from attacks in the USA (20 attacks), Germany (1), Canada (2), Brazil (1), and Finland (1). The average age of the attacker was 25.32 years, with the youngest being 12 and oldest, 53 years of age. There was a range of weapons used by the assailants, and they were mainly guns (23), but also aircraft (2), a smoke bomb (1), and a hammer (1). There was only one female assailant in the data, and 19 of the attacks occurred inside of schools (see Table 58 for suicide attacker modus operandi and summary statistics). The anonymous data was sourced from random news reports, web articles, personal blog posts, book extracts, and an oration transcript from healthy people still alive today. See Table 58 for a list of the 25 suicide attacker modus operandi and summary statistics. A total of 60 records were used in this series of experiments to characterise a suicide attacker.

Using the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker *et al.*, 2015), Egnoto and Griffin (2016) identified spree killers' negative emotions, and anger vocabulary use was significantly higher than students who did not commit suicide. Following the approach by Egnoto and Griffin (2016), we use LIWC to extract seven sentiment tags (emotional tone, affective process, positive emotion, negative emotion, anxiety, anger, and sadness) from our dataset. Mann-Whitney U testing was then conducted to examine if anger and negative emotion can statistically separate the data.

We assess the results using the LIWC sentiment tags and compare the results from our dataset to the Egnoto and Griffin (2016) results. We then extended their approach to a rules-based scoring clarifier using our data. This approach is continued using RPAS. We used step-wise multiple regression analysis to train the model and determine if it is possible to identify suicide attackers in cyberspace better than using LIWC. The resultant regression analysis unstandardized coefficients were used to weight the data, and the weighted RPAS variables were summed for a single score and plotted. A Jackknife Estimation Method (McIntosh, 2016) variation of 5-fold cross-validation was conducted with 12 random posts in each fold, to determine the optimum regression score that most accurately differentiated the RPAS data.

Here we assume that a suicide attacker could be suffering from depression and stresses from life change events (Corner & Gill, 2015). We attempt to better classify the model and added additional data containing non-suicidal blogger posts whose writing contains linguistic markers for depression (see Chapters 8 and 9 for the 45 Iris Murdoch and P.D. James data). The LIWC sentiment tag approach is also used to see if the classification can be improved and can better separate people with depression from suicide attackers, who may well suffer similar life change stressors.

A 65,800-word sample is achieved through 33,111 words from 25 suicide attacker notes and manifestos and 32,689 words from 35 anonymized blog posts and online articles from still living people on the internet.

The data is pre-processed using the Stanford Parts Of Speech (POS) Tagger (Toutanova & Manning, 2000) to remove all punctuation, numbers, and symbols before creating a nine-element array using RPAS to conduct step-wise multiple regression analysis.

Initially, the data is randomised, and 50 records are used to classify the data using stepwise multiple regression analysis. Five suicide attackers and five normal, non-suicidal blogger's records are held aside. All 60 records are then used to plot RPAS.

The model validated with a Jackknife (McIntosh, 2016) variation of 5-fold cross-validation with 12 samples in each of the five folds. This is done to determine the average classification value to fine tune the classifier before plotting Receiver Operating Characteristic (ROC) Curves are calculated.

We further draw on earlier studies (see Chapters 8 and 9), where two sets of RPAS data from Iris Murdoch and P.D James exist. One is 45 sample of both authors (each of size 4,000 words), and another 104 sample from Iris Murdoch only (each of size 1,000 word.)

The data is processed to create a signature using RPAS for each suicide note, manifesto and blog post (for a more detailed explanation, refer to the Methods Section 3.4.2). Details on The Linguistic Inquiry and Word Count Tool (LIWC) can be found in the Methods Section 3.6. The approach taken with the Mann-Whitney U-Test can be found in the Methods Section 3.7. Here, we use the Mann-Whitney U-Test to compare the two groups across each of the seven sentiment tags from LIWC. Stepwise multiple regression analysis is conducted, and details can be found in the Methods Section 3.8. Stepwise multiple regression analysis is conducted to construct a model of the data that classified a suicide attacker.

K-fold cross-validation analysis is detailed in the Methods Section 3.9. By setting *k* to 5, 60 samples are split into five groups containing 12 randomly assigned samples. Conducting multiple regression analysis five times using RPAV as the independent variables, the model is trained each time with a different fold left out so that there are five sets of unstandardized regression coefficients. The regression scores are calculated for the five folds, and the resultant accuracy scores are compared between a regression score range of 0.5 - 3.0. The optimum regression score value to classify the data is determined, and an averaged accuracy of the technique is calculated.

10.3 Analysis

10.3.1 Testing on LIWC Emotions

To ensure that the suicide attacker data is similar to spree killers' and contains significantly higher negative emotions and anger vocabulary than a normal person's, seven sentiment tags are extracted (emotional tone, affective process, positive emotion, negative emotion, anxiety, anger, and sadness) using the Linguistic Inquiry and Word Count tool. Mann-Whitney U testing (Section 3.7) is conducted to examine if anger and

negative emotions can statistically separate a suicide attackers' note or manifesto from a normal blog post (see Table 29). We adopt the common practice of ignoring the alternative Wilcoxon W and the Z scores given the small sample size and focus predominantly on the Asymptotic Significance (two-tailed) P values whose rankings are reflected in the Mann-Whitney U scores. The tests found statistically significant differences in anger ($p < .001$), negative emotion ($p = .002$), emotional tone ($p = .01$) and affective process ($p = .048$). Mann-Whitney U tests results highlighted anger as the most statistically significant differentiator between suicide attackers and non-suicidal people ($U = 151.5$, $p < 0.001$), followed by the use of negative emotion ($U = 219$, $p = 0.002$).

Even though our data is different, the most significant two sentiments, anger, and negative emotion, correlate with the findings in Egnoto and Griffin (2016). In Figure 28 we can separate the data with an anger value of 0.85 and a negative emotion value of 2.5. The majority of the non-suicide articles have low anger (74%) and low negative emotion values (71%), while the suicide attackers demonstrate high levels of anger (95%) and high levels of negative emotion (85%). In Figure 28, anger is a better differentiator between both groups, but there is some ambiguity between 0.85 – 1.5. At 0.85, 3 (5%) of attackers would be incorrectly classified with 9 (15%) of normal bloggers being incorrectly classified.

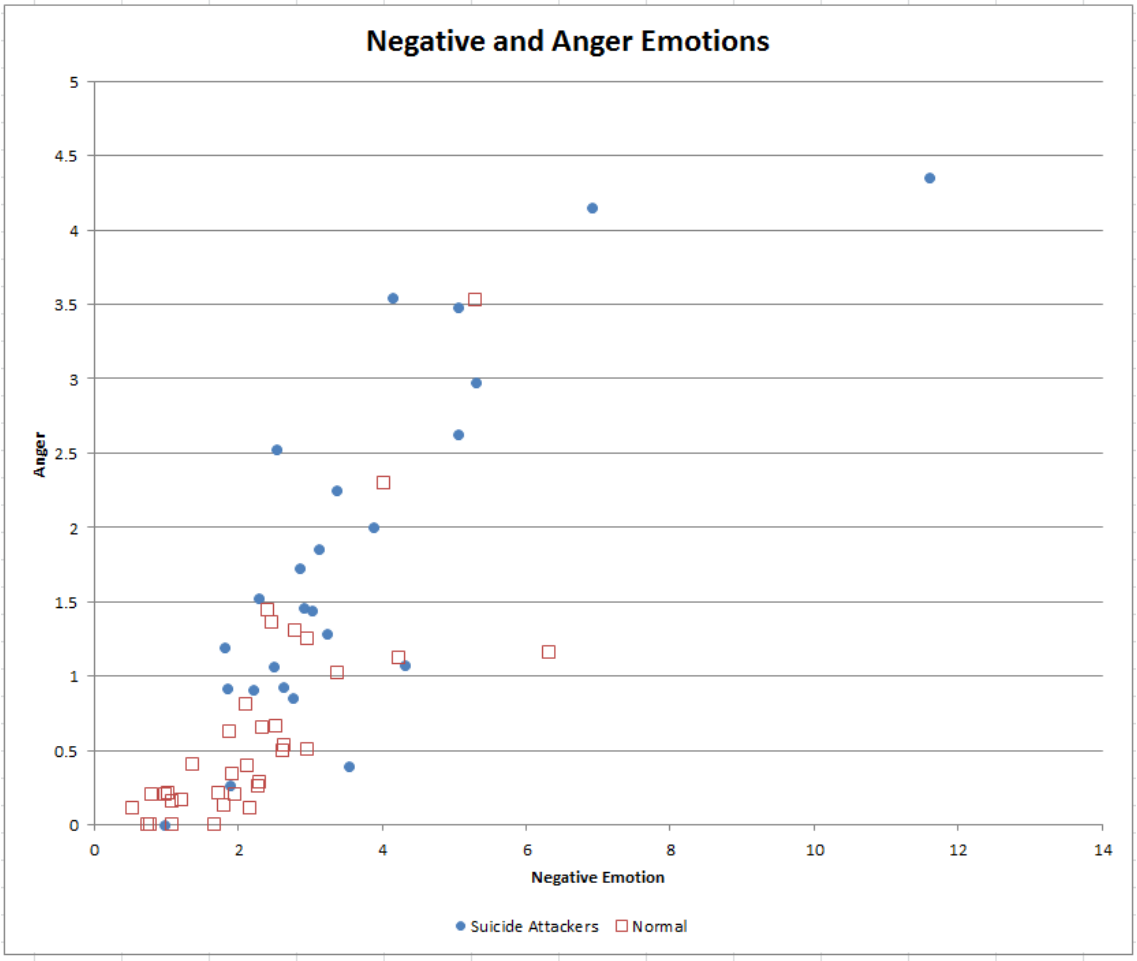
Table 29: By examining the Asymptotic Significance (two-tailed) p-values, the seven linguistic emotion categories from LIWC2015 highlight statistically significant differences ($p < .05$) between suicide attackers and normal blog posters in the area of anger, negative emotion, emotional tone, and affective process. As can be seen, Anger followed by Negative Emotion is the most significant (p values <0.01) and have the two smallest Mann-Whitney U scores. Due to the small sample size, we ignore the alternate Wilcoxon W and Z scores as is common practice.

Test Statistics[a]

	Tone	Affect	Positive Emotion	Negative Emotion	Anxiety	Anger	Sadness
Mann-Whitney U	257.500	296.000	346.000	219.000	335.500	151.500	419.000
Wilcoxon W	582.500	891.000	671.000	814.000	660.500	746.500	744.000
Z	-2.570	-1.979	-1.212	-3.160	-1.380	-4.196	-.092
Asymp. Sig. (2-tailed)	.010	.048	.226	.002	.167	.000	.927

a. Grouping Variable: Suicide Attacker r Normal

Figure 28: As expected from the LIWC negative emotion and anger categories, the majority (74%) of the non-suicide articles have low anger and negative emotion values, and 9 of them are away from the main group (anger > 0.85). All but 3 of the suicide attacker data is higher than 0.85. The results of Mann-Whitney U tests highlighted anger as the most statistically significant differentiator between suicide attackers and non-suicidal people (U=151.5, p < 0.001), followed by the use of negative emotion (U=219, p=0.002). In this diagram, anger is a better differentiator between both groups, but there is some ambiguity between 0.85 – 1.5. One suicide attacker has been omitted here because its value was -1.16615 and a large false negative.



It is clear that the underlying sentiment in the data, that there are differences between both groups. Receiver Operating Characterisation (ROC) Curves (Section 3.16) were calculated for both Anger and Negative Emotion (Figure 29), to take into account false positive rate (Specificity) and true positive rate (Sensitivity). The Area Under the Curve (AUC), Table 30, demonstrates that Anger provides a better overall classification than Negative Emotion (81.1% versus 73.7%).

Figure 29: ROC curves for Anger and Negative Emotion. Here Anger tracks the left-hand border and then the top border at a better rate than Negative Emotion suggesting Anger is a more accurate test. Here we can see the differences between Sensitivity and Specificity, showing that Anger has less false positives (Specificity) and Anger has more true positives (Sensitivity) than Negative Emotion

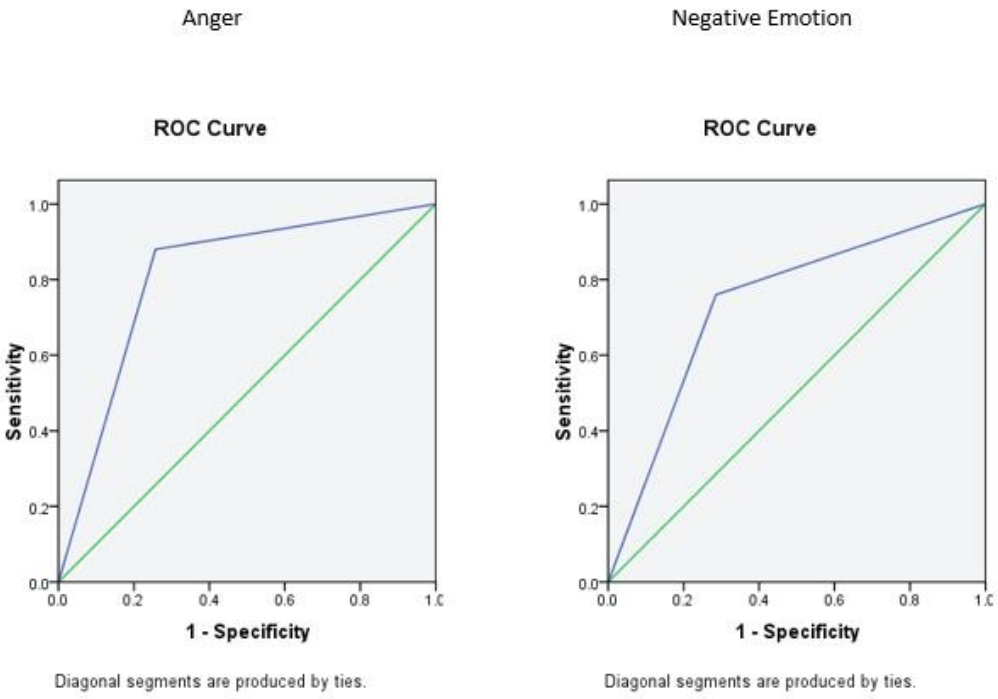


Table 30: The results of the Area under the curve (AUC) for anger and negative emotion showing the better ROC curve classification rates for anger over negative emotion.

Area Under the Curve

Test Result Variable(s):Anger and Negative Emotion

Variable	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
Anger	0.811	0.058	0	0.697	0.926
Negative Emotion	0.737	0.067	0.002	0.606	0.868

The test result variable(s): Anger and Negative Emotion have at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

10.3.2 Testing using RPAS

After having seen how successful LIWC sentiment emotions can be to separate suicide attackers from normal bloggers, we use **RPAS** to ascertain its usefulness as a classifier. From RPAS, an eight-element vector is constructed of the 60 items. We conduct step-

wise multiple regression analysis using SPSS (Chapman, 2017). We use 50 of the dataset records (keeping 5 suicide attackers and 5 normal non-suicidal bloggers aside) and train the model, by removing one RPAS (including VAHOG) variable at a time until only statistically significant variables remain. When these four variables are combined, the synergy between the variables produces a statistically significant result to predict suicide attackers ($F(4, 46) = 11.152$, $p < .0001$, $R^2 = .492$). These are the Richness (R), Personal Pronouns (P), Referential Activity Power (A), and the Visual (V) variable from the Sensory Adjectives (S) category. These values are identified as RPAV.

Using the resultant unstandardized coefficients and the constant variable regression score from the regression analysis, each of the RPAV variables is multiplied by its regression coefficient and sum the constant variable to the results. All 60 data have regression scores applied and are plotted (Figure 30). We find a regression score of 1.5 separates suicide attackers and normal blog posters with 83% accuracy. Seven (~11.6%) of attackers would be incorrectly classified, while 4 (~6.6%) of normal bloggers are incorrectly classified. Using a range between 1.5 – 1.7 captures most of the false positives and negatives, and outside of this range the classifications are correct except three. However, this introduces ambiguity.

To fine tune the RPAV technique, the data is randomized, and 5-fold cross-validation is conducted with 12 random posts in each fold to determine the optimum regression score that most accurately classifies the data. The results are plotted below (Figure 31) and highlight an average accuracy of 81% (78.3 - 85). It indicates a regression score of 1.45 will provide the best classification.

Receiver Operating Characterisation (ROC) Curves (Fawcett, 2006) were calculated for RPAV using the regression score of 1.45 to take into account false positive rate (Specificity) and true positive rate (Sensitivity) and compared with the LIWC anger values (Figure 32). We see that the LIWC Anger category has a higher number of true positives, while RPAS has a lower number of false positives. The Area Under the Curve (AUC) Figure 32, demonstrates that RPAV (RPAS with only the Visual modality present) provides a slightly better overall classification than Anger Emotion (83.1% versus 81.1%).

Figure 30: Stepwise Multiple Regression using the four RPAV elements on the 25 suicide notes and manifestos from suicide attackers and 35 'normal' blog posts and articles from people who are not mass attackers or have committed suicide. This method is better at indicating false positives, but not as effective with false negatives.

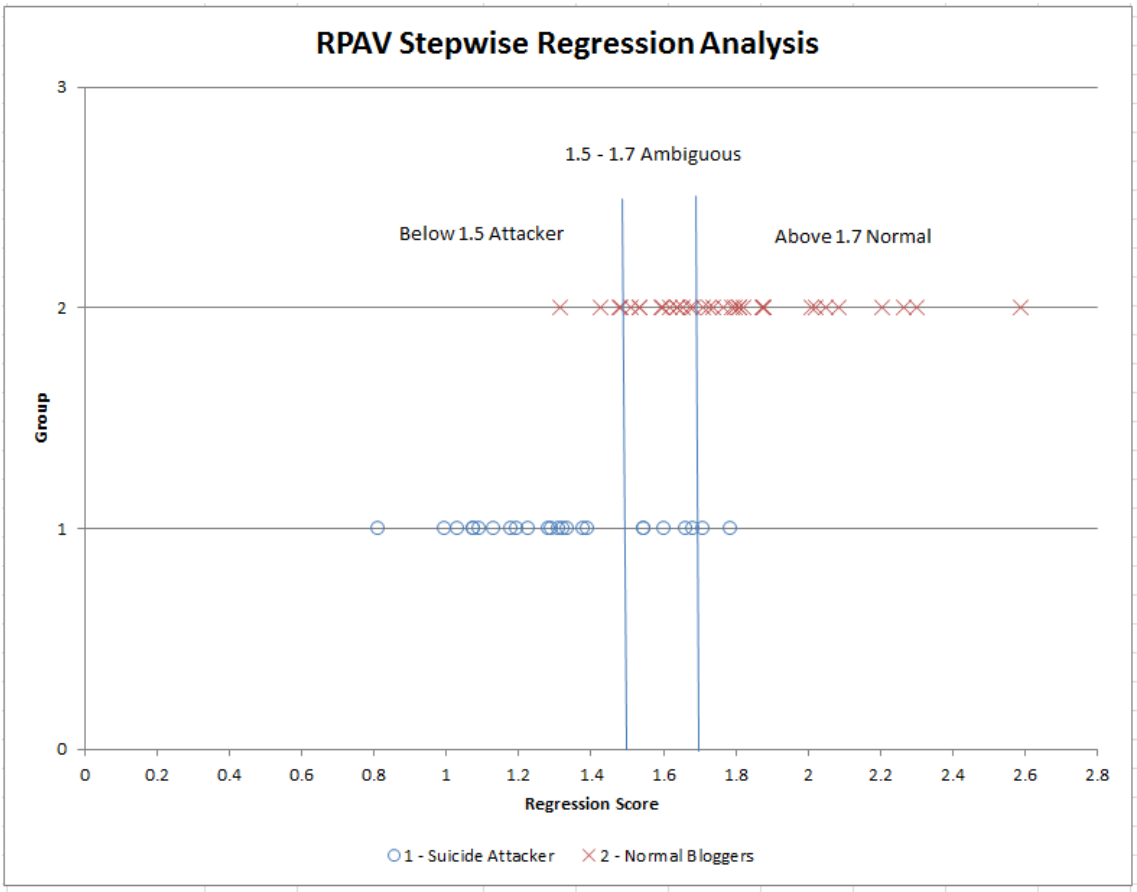


Figure 31: The original classification accuracy for regression scores between 0.65 - 2.5 is overlaid against the results of 5-fold cross validation showing the minimum and maximum ranges. As can be seen, the highest accuracy is achieved with a score of 1.45

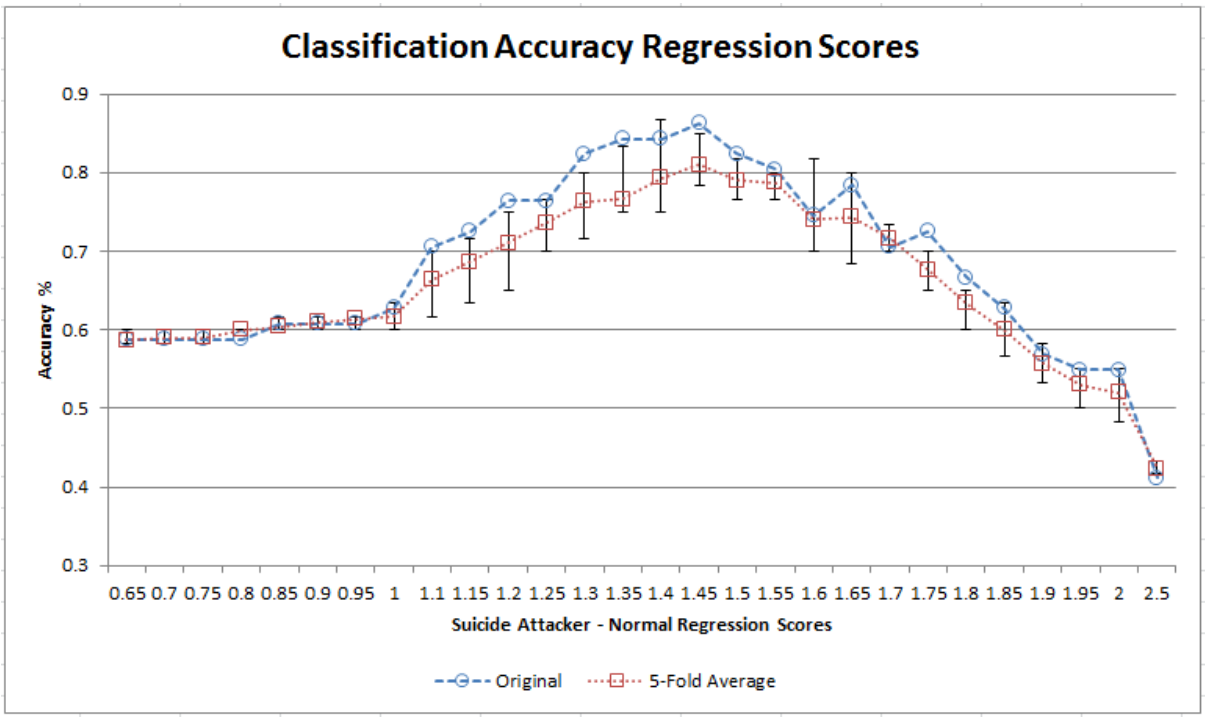


Figure 32: ROC curves for Anger and RPAV. Here RPAV tracks the left-hand border better than Anger and not quite as good as Anger on the top border suggesting these two tests are accurate. Here we can see the differences between Sensitivity and Specificity, showing that RPAS has less false positives (Specificity) and Anger has more true positives (Sensitivity).

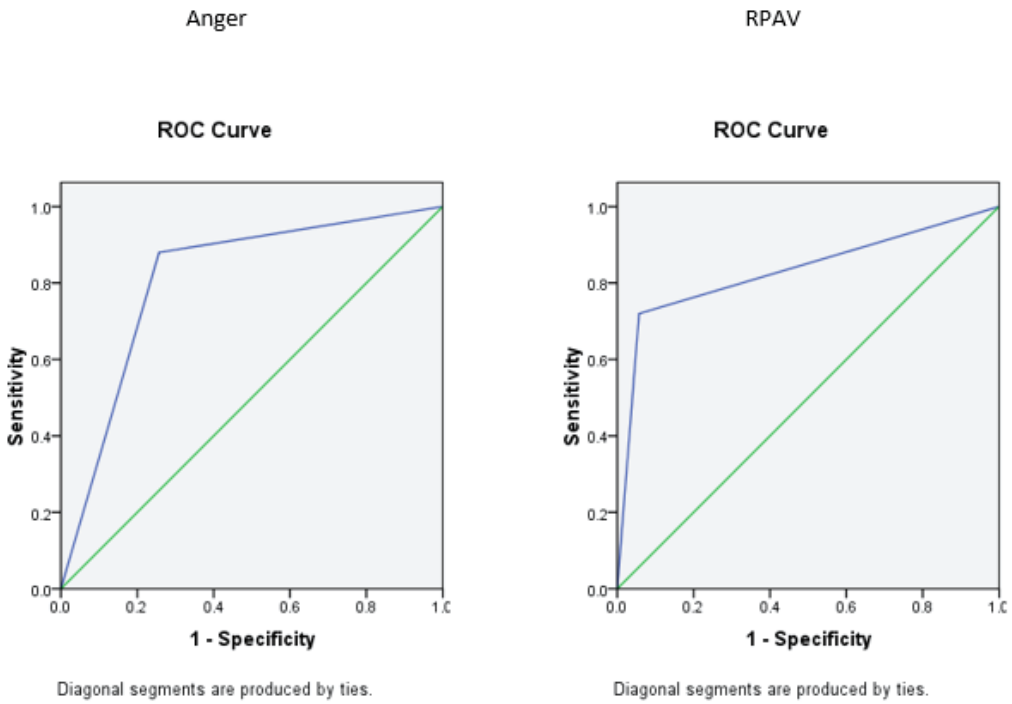


Table 31: The results of the Area under the curve (AUC) for anger and RPAV showing the better ROC curve classification rates for RPAV over anger.

Area Under the Curve

Test Result Variable(s):Anger and RPAV

Variable	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
Anger	0.811	0.058	0	0.697	0.926
RPAV	0.831	0.059	0	0.715	0.948

The test result variable(s): Anger and RPAV have at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

10.3.3 Testing for depression

Having verified the model, an additional 45 samples (Iris Murdoch and P.D. James) are added to the data to determine if it is possible to identify sentiment tags that highlight differences in Iris Murdoch because of her known depression. Although the data includes works where Murdoch's depression and Alzheimer's disease, Mann-Whitney

U testing (Section 3.7) of the LIWC data (Table 32) highlights Iris Murdoch is no different to normal bloggers, in that the writing is significantly different from suicide attackers in emotional tone (U=161, p=0.017), affective process (U=153.5, p=0.010), negative emotion (U=130, p=0.002), anger (U=69.5, p < 0.001). Excluding tone, P.D. James is little different (affective process (U=96.5, p=0.004), negative emotion (U=122.5, p=0.036), and anger (U=49.5, p < 0.001)). However, what is different is that both are significantly different from suicide attackers in anxiety (Murdoch - U=158, p=0.014, and James - U=89.5, p=0.002). Comparing Iris Murdoch and P.D. James to the normal bloggers, there are no significant differences with any of the sentiment tags in the case of Iris Murdoch, but P.D James is significantly different in the areas of positive emotion (U=154.5, p=0.007), and anxiety (U=174.5, p=0.023). A comparison between Iris Murdoch and P.D. James shows significant differences in tone (U=139, p=0.013), and positive emotion (U=123.5, p=0.005).

Table 32: Mann Whitney U-Test results of Iris Murdoch and P.D. James compared to Suicide Attackers and Normal non-suicidal bloggers. Note: due to the small sample size, we ignore the alternate Wilcoxon W and Z scores as is common practice.

Test Statistics[a]

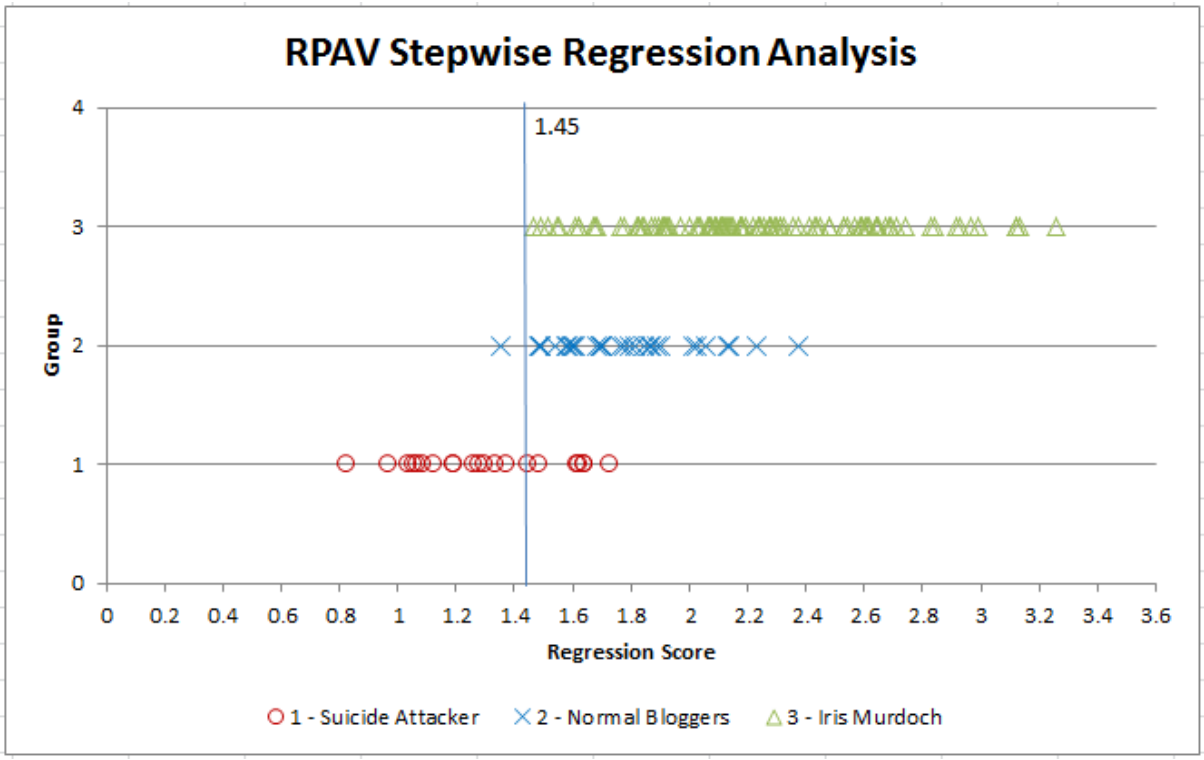
	Rank of Tone	Rank of affect	Rank of posemo	Rank of negemo	Rank of anx	Rank of anger	Rank of sad
Iris Murdoch vs Suicide Attackers							
Mann-Whitney U	161	153	223.5	130	158	69.5	242.5
Wilcoxon W	392	504	454.5	481	389	420.5	473.5
Z	-2.397	-2.568	-1.059	-3.06	-2.467	-4.356	-0.653
Asymp. Sig. (2-tailed)	0.017	0.01	0.289	0.002	0.014	0	0.514
P.D. James vs Suicide Attackers							
Mann-Whitney U	163	96.5	180	122.5	89.5	49.5	173.5
Wilcoxon W	394	286.5	370	312.5	320.5	239.5	404.5
Z	-0.989	-2.79	-0.528	-2.086	-2.992	-4.063	-0.705
Asymp. Sig. (2-tailed)	0.333	0.004	0.611	0.036	0.002	0	0.486
Iris Murdoch & P.D. James vs Normal non-suicidal bloggers							
Mann-Whitney U	190	220.5	154.5	259.5	174.5	277.5	227
Wilcoxon W	380	410.5	344.5	724.5	639.5	742.5	692
Z	-1.949	-1.324	-2.678	-0.523	-2.268	-0.154	-1.19
Asymp. Sig. (2-tailed)	0.051	0.186	0.007	0.601	0.023	0.878	0.234

a. Not corrected for ties.

Having verified the model, an additional 104 samples of Iris Murdoch are added to the data (all 1000 words in size) to determine if it is possible to separate Iris Murdoch from a suicide attacker using the RPAV technique. These are compared these against

normal bloggers and suicide attackers Figure 33 and the results highlight that Iris Murdoch identifies as normal with regression scores above 1.45.

Figure 33: The RPAV stepwise regression results, this time showing the addition of one hundred and four 1,000 word samples of the works of Iris Murdoch, which all occur above the classification value of 1.45.



All of the data processed using the Linguistic Inquiry and Word Count Tool (LIWC) extracted seven different sentiment tags (emotional tone, affective process, positive emotion, negative emotion, anxiety, anger, and sadness). Mann-Whitney U testing confirmed that anger and negative emotions were both statistically significant and could be used to separate a suicide attacker's writing from normal posts, as per Egnoto and Griffin's (2016) spree killers study. While the Mann-Whitney U testing highlighted four statistically significant sentiment tags in anger ($p > .001$), negative emotion ($p = .002$), emotional tone ($p = .01$) and affective process ($p = .048$), visually, only anger and negative emotion provided the best results with anger at a value of 0.85 providing the most accurate separation (76.5%) as seen in Figure 33. Of the three suicide attacker scores below this anger value of 0.85, no other similarities could be found by examining the six other sentiment tags produced by LIWC. Dutiel, the only female, (scored 0) de Oliveira (0.27), and Morrison (0.4) had low anger scores, and she showed no other obvious similarities in age, victim count, locality or year of the event that might categorise her as a suicide attacker.

Until now, RPAS ability to separate broad styles of text in one group of authors from another was untested. By constructing a model using eight variables in RPAS (including the five sensory variables, VAHOG), step-wise multiple regression analysis highlighted Richness (R), Personal Pronouns (P), Referential Activity Power (A), and the Visual (V), or RPAV, as statistically significant to predict suicide attackers ($F(4, 46) = 11.152$, $p < .0001$, $R^2 = .492$).

From RPAV, the resultant unstandardized coefficients were used to classify the data regression scores (see Figure 30) at using a value of 1.5 with an accuracy of 86%, and using a range between 1.5 – 1.7 didn't improve the results. However, compared to the LIWC method using negative and angry emotions (76.5% accuracy), it performed better as a classification technique.

A validated RPAV model using 60 randomised samples with 5-fold cross-validation (Figure 31), highlighted an optimum regression score of 1.45, with an average accuracy of 81% (78.3 - 85).

Through testing, limitations on the file size were discovered. Christopher Dorner's manifesto was almost 13,000 words, with a resultant regression score of -1.5. While he still classified correctly, the regression score was very distant from the smaller suicide notes. The algorithm is suitable for file sizes of up to around 2000 words, but it has been optimized for files up to 800-1000 words. It will work on file sizes around 4000 words, and upwards using a different set of unstandardized coefficients that use all five of the sensory element, not just the Visual variable. We have not reported on this approach because it is inferior to the approach we have identified above. As an alternative to addressing the larger file size issue, the Referential Activity Power variable could be recalculated so that the results are then square rooted (effectively removing the original squaring factor). This would reduce the range of the large negative scores observed. However, it would not change the existing relationship between the findings of the different authors.

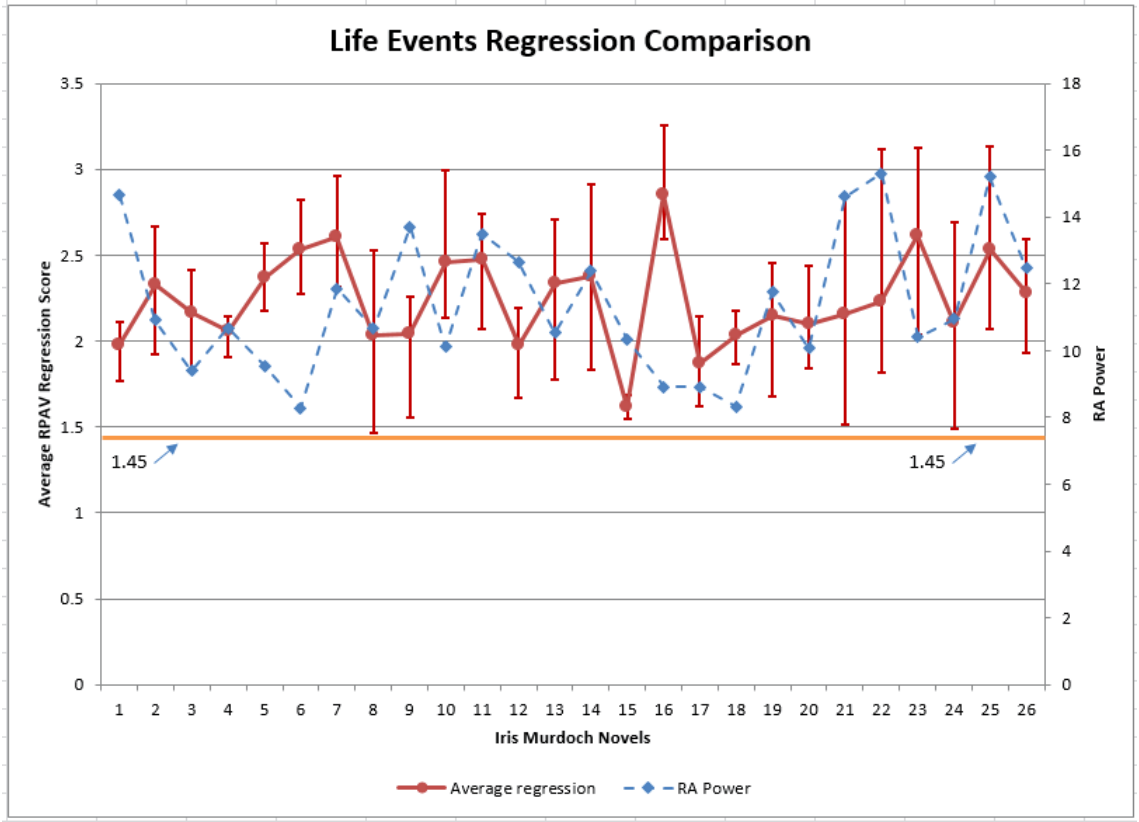
10.3.4 Separating depression from suicide attackers

There were two key goals for this chapter. One was to determine if it were possible to separate a suicide attacker's notes and manifestos from a normal blogger, and the second was to examine if the algorithm could further separate a person with depression (Iris Murdoch) from a suicide attacker. In previous research on Iris Murdoch (Chapter 8 and 9), data with known markers for depression had been used to

help identify possible tipping points that might indicate self-radicalisation. Retesting the Murdoch data with RPAV would refine the wider study.

Mann-Whitney U testing (Section 3.7) of Iris Murdoch was unable to highlight significant differences in sentiment against the suicide attacker writing outside of anxiety. Comparing Iris Murdoch and P.D. James to the normal bloggers, there were no significant differences with any of the sentiment tags in the case of Iris Murdoch, but P.D James was significantly different in the areas of positive emotion and anxiety.

Figure 34: A comparison between the known depressive periods in Iris Murdoch's life (novels 3, 5, 6, 16, 17, and 18) where the RA Power levels fall below 10. These are compared to the stepwise regression analysis results. Note, there is no correlation between these low regression scores where they approach a value of 1.45 – the separator between a suicide attacker and a normal blogger – which suggests it might be possible to separate depression in a normal person from depression in a suicide attacker.



Using the LIWC sentiment tags, the results of the Mann-Whitney U test could not statistically separate Iris Murdoch from a suicide attacker. Therefore, one hundred and four 1,000 word segments were selected from the 26 novels of Iris Murdoch (4,000 words per novel). This data was compared to normal bloggers and suicide attackers (Table 58) and found it classified correctly above 1.45. It was known that Iris Murdoch's novels contained signs of depression or where she had suffered from life events (see Chapters 8 and 9) as shown in the RPAS Referential Activity Power (A) variable falling below a value of 10 (novels 3, 5, 6, 16, 17, and 18). By taking the 104

1,000 word samples of Iris Murdoch and averaging each of the 4 samples from the 26 novels (and including the minimum and maximum scores for each), the results of the regression analysis was compared to the known signs of depression. While none of Iris Murdoch's work identified as a suicide attacker, there were five instances where they fell at or below 1.55 (8 – 1.46, 9 – 1.55, 15 – 1.55, 21 - 1.51, and 24 - 1.48). These known points of depression from life events were compared to the five low regression scores. If the points correlated, it would have meant that depression was likely linked to the RPAV model. However, this was not the case. This finding suggests that RPAV is better than LIWC given it should be possible to separate general depression in an individual from the writing style of a suicide attacker, but more examples are needed.

10.3.5 Limitations

While this approach to the identification of a suicide attacker shows merit, it is dependent upon a longitudinal study of only two authors that highlighted depression from life events in Iris Murdoch. However, Murdoch's depression and anxiety have been well documented through her prolific habit of writing about herself throughout her life (Dooley & Nerlich, 2014; Martin & Rowe, 2010; Murdoch, 2015; Wilson, 2004), and there have been a number of scientific studies that verify it. In this chapter, her anxiety was also observed in the LIWC anxiety tag. With regards to the suicide attacker data, there is only one record for each attacker. While they would have been suffering from life events at that point in time, there is no way to compare their current state of mind to earlier times before they had become radicalised and make a proper assessment of their likely levels of depression at the time they wrote their manifestos and final suicide notes.

10.4 Conclusion

In this study of suicide notes, final manifestos, book extracts, newspaper articles, blog posts, and orations, a clear differentiation between the writing style of suicide attackers' final manifestos and suicide notes was found when compared to the posts of normal bloggers. While this approach lends itself to be automated based on a resultant regression score of 1.45, it might be useful in separating the depressed posts of normal people from suicide attackers. Importantly, the findings indicated no correlation between a normal blogger's low periods in their life and their low regression scores. The authors suggest that this exploratory approach using step-wise multiple regression

analysis and the four RPAV variables has the potential to identify suicide attackers in cyberspace.

10.5 Summary

In this chapter, the study of the final manifestos and notes of suicide attackers, we have been able to improve the classification of a terrorist's theoretical writing, and separate their stylistic signature from both a normal person and somebody with depression who suffered from life changes see if we can separate their writing from normal blog posts. The most significant findings from this study were the RPAV method was more effective than the LIWC technique that used the anger category or the negative emotion category. Using RPAS, we were also able to separate the writing of Iris Murdoch, someone with depression and suffering from life events, from that of suicide attackers who were likely to have been suffering from critical life events, and possibly also depression.

Discussion and Conclusions

11.1 Introduction

In this chapter, the findings of the thesis are discussed. We summarise the aims and scope of the research thesis and the key issues that have emerged from the seven studies that were conducted in a three-phased approach to the self-radicalisation problem. We discuss our contributions to the theory and the limitations of those findings. Conclusions are drawn from the data presented in the studies with suggestions made to extend the scope and direction for future work.

Our focus in this research thesis is on the stylometric processing of sensory open source data. The aim is to create a method that extracts key linguistic features, or attributes, from a person's writing style or speech that can characterise self for identification and be used to predict self-radicalisation is the principal focus for this research thesis. We use biomarkers for personality that are reflected in language to improve authorship profiling, examine changes in time from normal aging and from disease and depression. We use this to create a mathematical model of identity that can show tipping points in a person's state of mind. However, we still have to define an input and output framework, so the model was more complete.

This research is driven by the terrorism problem facing Australia, from the growing threat of returning foreign fighters from Iraq and Syria, its increase within Australia's wider near-region, specifically Indonesia and the Philippines, and from home-grown threats. Research on terrorism is underdeveloped, and no fully scientific theory can explain these phenomena. Its significance can be seen by the rise within our region of planned attacks, both successful and foiled where many attackers were known to authorities, and existing risk assessment systems had failed. The anonymity of social media and the internet makes it difficult to identify who is becoming radicalised and when a person with radical or politically motivated beliefs changes to act in a violent way.

The program of research reported in this thesis focuses on the lack of authorship analysis tools and the identification of anonymous authors to identify people that might be insurgents, 'insiders', or a lone wolf. It draws on neuropsychology and neuroscience markers within the brain to characterise an author's identity using features, or markers within writing to identify self. By looking at cues, with CSD (Section 3.4.2), we posit the point prior to when a lone wolf commits a terrorist act might be identifiable to provide an early warning and prevent a possible crisis.

Drawing on the four research questions (Section 1.3), the objective of this research is to separate the identity of individuals and to highlight changes within them that indicate self-radicalisation. This research shows that using personality for identity (RPAS) it is possible to create a personal signature of individuals (research hypothesis 1) and to separate 'normal' writing from that written before a terrorist attack (research hypothesis 4). When taking into account the 'normal' changes in a person's signature over time (research hypothesis 2 and 3), it is possible to use techniques to visualise CSD (Section 3.4.2) and determine tipping points of change in an individual (research hypothesis 3).

Access to a larger and varied set of real-world data permitting, using the techniques from this research on disenchanted people, it might be possible to identify the tipping point before a terrorist become self-radicalised and stop lone wolves before they act.

11.2 Significant and Original Outcomes

There were a number of significant outcomes from the thesis. We developed a new neuro-linguistic technique, RPAS, which creates a signature of an individual's identity from their writing mapping personality or self.

We found that a Multi-disciplinary approach is quite effective to characterise personality from writing. While there were a number of statistical analysis techniques used in this thesis to visualise identity (AR1, G1, LDA, PCA, Seriation, VSM), RPAS itself stems from a number of disciplines.

Richness (R) is grounded in ecology, and studies of species diversity and species density. It is linked to neuroscience and studies into age and the decline of cognitive functions of the brain. It is also used in an inverted form as the Type Token Ratio (TTR) in computational linguistic studies. Personal Pronouns (P) is grounded in computational linguistic studies of gender identity. Referential Activity Power (A) is

grounded in neuropsychology and used in clinical studies of depression. Sensory Adjectives (S) and the five modalities (visual, auditory, haptic, olfactory, and gustatory) are grounded in neuroscience and the function of the sensory cortex. They also have their basis in Neuro-Linguistic Programming (NLP).

We found that identity does change over time in an individual due to natural aging. While healthily aging adults may also experience a decline in their cognitive abilities, it is seen in their language but is significantly less severe (Maxim & Bryan, 1994). This was reinforced through study 4 and 5 (Chapters 7-8), and demonstrated through a number of different techniques. Such as, from the way the RPAS signature and the use of seriation with noise technique was able to highlight the subtle characteristics of writing in Shakespeare's *Sonnets*.

When comparing the works of Iris Murdoch with P.D. James, a healthy aging control whose novels follow the patterns expected for normal aging elders (Le *et al.*, 2011), the signs of normal aging in P.D. James were demonstrated through falling sensory adjectives across all five sensory modalities (visual, auditory, haptic, olfactory, and gustatory). There were no statistically significant changes in function word to content word ratios throughout her writing career over 50 years. There were no statistically significant changes in Richness scores in the last 12 years of her writing before her death when compared to the previous 40 years of her writing career.

While autocorrelation is growing in interest within longitudinal psychology studies (Bringmann *et al.*, 2017) our approach is new. While mostly untested on writing, CSD with RPAS can show a tipping point. This was reinforced through study 6 (Chapter 9), and demonstrated by the use of the modified 1-lag autocorrelation (AR1) and Fischer-Pearson coefficient of skewness (G1) techniques on the RPAS Sensory variables. Signals observed in the AR1 and G1 techniques indicated tipping points and in Murdoch's case, an unusual change in her 15th novel. This was strongest using the combined sensory modalities of RPAS and closely followed the visual modality, which is a dominant modality (Posner *et al.*, 1976). This correlates with the findings of the previous study (Chapter 8 study 5) using Parts of Speech (POS) Function to Content word ratios and Richness. This could indicate the early signs of Alzheimer's disease at a point 20 years prior to Murdoch's final novel when her cognitive decline had become apparent. To support this claim, we provide an overview of AD, and its impacts on sensory processing in memory recall and creating new ideas.

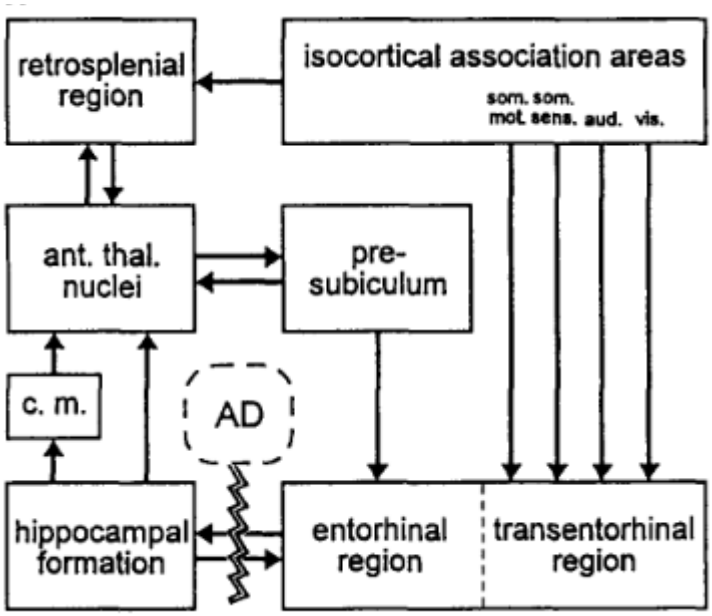
The Relationship between the progression of AD and the impacts of it on memory forming, recall, and sensory words is important at this point, as it reinforces all of the experiments conducted. There are six (Braak) stages in the development of AD. In their book on neurodegeneration of the cerebral cortex, Cechetto and Weishaupt (2016: pp84-92) provide a good explanation of these stages, summarised below. In the first two stages, the preclinical stages where there is an absence of clinical signs, neurofibrillary tangles (NFTs) develop in the transentorhinal cortex, and entorhinal cortex (EC) region. In stage three and four, the limbic stages or the pre-dementia phase where mild cognitive impairment (MCI) occurs, NFTs spread to the subiculum, the Cornu Ammonis (CA)1, and CA3 regions of the hippocampus (associated with memory and in particular long-term memory, and a part of the limbic system that regulates emotion) before reaching the dentate gyrus (DG), a region in the hippocampus responsible for the formation of new episodic memories. Plaques (abnormal clusters of a protein known as beta-amyloid) appear in the brain between the EC and the hippocampus. Patients experience mild memory loss and executive function decline (such as poor attention, manipulation skills, and a lack of ability to regulate behavioural control). Finally, in stages five and six, the neocortical stage where dementia is apparent, the NFTs severely affect the neocortical association areas and spread to the structures in the temporal, parietal and frontal lobes affecting the motor and sensory areas in the brain. In Figure 35, the pathways are shown in more detail (Cechetto & Weishaupt, 2016; Braak & Braak, 1995).

NFTs are composed of a number of smaller paired helical filaments (PHF), and the core protein in these filaments is the microtubule associated protein *tau* (Cechetto & Weishaupt, 2016: p84). Recent studies into AD show that prior to the formation of plaques and NFTs, there is a build-up of *tau* that collects in the densely-interconnected brain regions causing connections to weaken memory (Cope *et al.*, 2017).

The EC, where NFTs first form, acts as a central hub, a multimodal association area for sensory information from isocortical areas (providing inputs from visual, auditory, somatosensory, somatomotor, gustatory and olfactory cortices (Cechetto & Weishaupt, 2016: pp91-92). In the creation of new episodic memories, to encode our daily personal experiences, and to retrieve episodic memories, the outputs of neocortical association areas (sensory-specific temporal isocortical brain regions) are channelled through the EC to the hippocampus, and some of the earliest indicators of AD are when tests

highlighted a patient can recognise a visual object but has word-finding difficulty (Dickerson & Eichenbaum, 2010).

Figure 35: A map of the cerebral cortex showing the critical path of the entorhinal region, and where AD first occurs. Note: aud = auditory, ant thal nuclei = anterior thalamic nuclei, cm = mammillary body, som sens = somatosensory, som mot = somatomotor, vis = visual.



Using RPAS, it is possible to observe stylometric markers from Alzheimer's disease extending back 20 years before a medical diagnosis, which is better than the results from Garrard *et al.* (2004), Le *et al.* (2011), and Hurst and Feng (2012). We also observed depression in writing over time using RPAS. Some of the detail here has been mentioned in above, and reinforced through study 5 and 6 (Chapters 8-9), and demonstrated through times of low Referential Activity Power when Murdoch was known to be depressed. The Sensory element of RPAS highlights a lower use of Olfactory words which might be a marker for AD and cognitive decline. Referring back to our discussion on the progression of AD, the NFTs did not impact the primary sensory cortices at Braak's stage three and four where MCI occurs (Cechetto & Weishaupt, 2016), the olfactory bulbs were damaged very early (i.e. Braak's stage one or before (Esiri & Wilcock, 1984). Olfactory deficits in AD and MCI are reliably observed in multiple olfactory domains, including odour detection threshold, identification, and recognition and olfaction loss is an early marker for AD and MCI (Quarmley *et al.*, 2017).

It is possible to separate suicide attackers writing from normal people, and people with depression who are not suicide attackers using RPAS. This was reinforced through study 7 (Chapters 10), and demonstrated through step-wise multiple regression of the

RPAS variables. Four of the eight variables were statistically significant to predict the writing of suicide attackers when compared to normal blog posts. They were RPAV: Richness (R), Personal Pronouns (P), Referential Activity Power (A), and the Visual (V) variable from the Sensory Adjectives (S) category.

11.3 Answering the Research Questions

To answer the thesis research question (*Can the automated extraction of key linguistic attributes from text-based data identify an author's personality, or self, and be used to predict self-radicalisation*?), four questions (see Section 1.3 and 1.4) expressed as hypotheses tests were tested, as follows:

Hypothesis H_1 (*The stylistic fingerprint of a person's personality – their personal signature – can reveal their 'identity' from their writing style*) can be demonstrated by the experiments conducted in Chapters 4-6, which highlighted the following:

- PtoR, AtoR, and StoR plots highlighted stylistic differences between Shakespeare, Marlowe, and Cary writing chunks, and separated the four generally accepted Shakespeare scenes from Kyd in *Edward III*.
- Word Accumulation Curves of Shakespeare and Marlowe highlighted that both author's unique word use was similar.
- The known authored works of Shakespeare, Marlowe, and Cary were able to be separated from the known contested works believed to be written by other playwrights and poets using Principal Component Analysis (PCA) on the RPAS signatures. It also highlighted two works, *The Phoenix and the Turtle*, and *Venus and Adonis*, which are believed to be written by Shakespeare, are stylistically different from his other works.
- The Haptic and Auditory Sensory elements of RPAS were able to differentiate the known authored works of Shakespeare, Marlowe, and Cary using Stepwise Linear Discriminant Analysis (LDA).
- The four generally accepted Shakespeare scenes in *Edward III* were more stylistically similar to Kyd's authorship than Shakespeare's when using Cosine and minmax similarity detection and the Vector Space Method (VSM). The Imposters Method with Marlowe's *Hero and Leander* reinforced these findings.
- Seriation was effective at separating and clustering the works of Shakespeare and Kyd in *Edward III*, and adding different levels of noise into the seriation matrix was effective at testing the weakest connections.

- Kyd likely wrote the four Shakespeare scenes in *Edward III*, and Shakespeare and Kyd were the likely authors of the play.
- The 12 anonymous poems in *The Passionate Pilgrim* consistently clustered eight of the nine known authored works using a PtoR plot, Principal Component Analysis, Linear Discriminant Analysis, and the Vector Space Method.
- One of the commonly accepted Barnfield poems within *The Passionate Pilgrim* was identified as Shakespeare's work.
- All 12 anonymous scenes in *The Passionate Pilgrim* were allocated authorship except one believed to be written by the poet, Thomas Deloney.

Hypothesis H_2 (*A person's 'identity' changes over time because of life events, such as trauma, depression, and disease*) can be demonstrated by the experiments conducted in Chapters 7-9, which highlighted the following:

- Seriation with RPAS was able to create the best grouping of the Dark Lady sonnets by using Richness (R), and a person's internal or socially constructed gender expressed as feminine or masculine (P). However, if the gender of the author was known, then those aspects could be discarded, and a configuration of RAS also was as effective.
- When comparing Iris Murdoch's earlier work to that 12 years prior to her formal diagnosis of Alzheimer's disease (AD), there were indications of lower Richness and a falling Function to Content word ratio which are linguistic markers for AD.
- With the exception of RPAS Olfactory words, there was an increase in Murdoch's use of RPAS sensory adjectives in the last 12 years of writing where James had a decreased use of all sensory word modalities, suggesting that a comparative lower use of Olfactory words might be a marker for AD and cognitive decline, consistent with current neuroscience theory.
- The RPAS Referential Activity Power for Murdoch's works shows a considerable amount of variation compared to James', and there are at least three excursions that are much lower than James during times of significant life events.

Hypothesis H_3 (*The application of techniques to visualise the critical slowing down phenomena can identify changes in a person's moods, or shifts from one state to another, that might indicate a tipping point for self-radicalisation*) can be demonstrated by the experiments conducted in Chapter 9, which highlighted the following:

- Using the RPAS Sensory element, we visualised CSD using the AR1 and G1 techniques. The results showed rising trends followed by falls that were indicative of CSD near a tipping point.
- An unusual signal was observed around Murdoch's 15th novel in the AR1 technique using RPAS Sensory Adjectives. This could indicate the early signs of Alzheimer's disease at a point 20 years prior to Murdoch's final novel when her cognitive decline had become apparent.

Hypothesis H_4 (*The final writings of suicide attackers can be separated from 'normal' bloggers*) can be demonstrated by the experiments conducted in Chapter 10, which highlighted the following:

- Discriminant Regression analysis of RPAS separated suicide attackers' notes and final manifestos from non-attacker blog posts.
- Negative emotion and anger using the Linguistic Inquiry and Word Count (LIWC) program separated suicide attackers' notes and final manifestos from non-attacker blog posts.

- The RPAS method outperformed the LIWC technique using the anger category (86% versus 80%) with one less suicide attacker being incorrectly classified.

11.4 Impact

Drawing on the significance of this study, we consider who will benefit. It is likely that there are three general areas. The first is those who seek to use an alternate approach to authorship identification that draws on personality, or self to separate an author's writing. The second is for those that see the benefits in using critical slowing down to develop early warning signs of tipping points from sensory data for the self-radicalisation problem prior to an attack. Third, the research might be useful in highlighting Alzheimer's disease 20 years prior to a formal medical diagnosis.

There were several unanticipated outcomes in the study. As far as we know, there is little to no research into mapping the individual sensory modalities (sight, hearing, feeling, smell, and taste) and using them for identity. The use of two-modality sensory adjectives in identity is novel. The surprises come from the analysis of the sensory aspects of Iris Murdoch, both in terms of the different Olfactory scores when comparing them against her other sensory modality scores and also the differences

against the P.D. James' scores, and in using Critical Slowing Down to highlight the tipping point on the amalgamated Sensory score.

Having seen that the writing in Iris Murdoch, a person who developed Alzheimer's disease, is different 12 years prior to her formal diagnosis, we find that with the exception of Olfactory, her sensory scores are also higher and not lower 12 years prior to diagnosis. Conducting techniques that visualise the critical slowing down phenomena, a change was predicted around the 12-year mark prior to her formal diagnosis with the disease, but the surprise came in seeing an unusual response in the data much earlier, about 20 years prior to her last novel where the linguistic signs of her disease were clear. While these are the results of a comparison between only two authors, they were unusual given the known, recorded aspects of her writing from other people's studies. This approach is new, but drawing on an ecosystems framework and CSD, the predictive early warning signals have been found in type II diabetes (Li *et al.*, 2013) and in clinical depression (van de Leemput *et al.*, 2014), with a view to using the approach for Alzheimer's disease (Hubin, *et al.*, 2016).

To support the sensory observations, we conduct Principal Component Analysis on the sensory VAHOG elements. We measure the total percentage of variance these elements contribute to the overall extracted components and iteratively remove and replace them to determine the impact that each one has on the total component's variation. We find that except for the Olfactory element, the results are relatively similar across V, A, H, and G for both Iris Murdoch and P.D. James. In the case of the Olfactory element, we see a large and significant contribution, and therefore a difference, which this element plays within the Iris Murdoch data. While Alzheimer's disease has been known to impact normal olfactory function with suggestions that olfactory loss may be a biomarker for AD and cognitive decline (Wesson *et al.*, 2010; Woodward *et al.*, 2015), it was a surprise to be able to observe a difference in olfactory sensory adjectives.

As far as we know, there is also little to no research in using the concept of Referential Activity from clinical studies in depression as a linguistic marker for identity. The use of highly concrete and imageability articles, pronouns, conjunctions, and prepositions were shown to relate to self, successfully predicting the authorship of Elizabethan playwrights. The surprises come from the analysis of the Referential Activity Power where Iris Murdoch's lowest scores correlated to difficult and depressed times in her life. Given that P.D. James writing began from a dark place, we allocated scores that fell below 10 as a significant event, of which P.D. James did not have.

Adding different levels of noise to the combinatorial technique seriation and test the strength of the connections between authored texts successfully predicted the authorship of the Elizabethan playwrights, and also highlighted subtle differences in single-authored works.

The use of modified variants of the 1-lag autocorrelation and Fischer-Pearson coefficient of skewness equations with the RPAS algorithm detected tipping points that highlighted Alzheimer's disease progression around 20 years before a formal medical diagnosis occurred.

11.5 Limitations

There are several limitations to these findings and the RPAS algorithm. The major limitation has to do with document size and is linked to the Richness element of RPAS. The type-token ratio (TTR) can be considered a variant of Menhinick's (1964) species diversity equation that measures vocabulary richness. While TTR is one of the oldest and easiest ways of measuring richness, it is dependent on text size, and while many attempts to reduce this problem have been proposed no one has been fully successful (Kubát, & Milička, 2013; Poiret & Liu, 2017). The biggest criticism of TTR is that it should not be used on its own, rather it should be incorporated it into a larger suite of techniques (Kubát & Milička, 2013; Vermeer, 2000). We avoid this by using the **RPAS** multivariate technique.

However, this study indicates that with large file sizes over 20,000 words, there are better-suited techniques, such Yule's K, or Rényi's higher-order entropy which perform well and are independent of text size (Kimura & Tanaka-Ishii, 2014; Tanaka-Ishii & Aihara, 2015). To alleviate this, a recommendation would be to chunk the data into smaller sized files of equal size, mark them as the same author, and highlight the centroid. A file size of 4000 words performed well. However, it is possible that another file size might perform better. We have used sizes as small as 80 words. In this study, keeping all data files a similar length when comparing results over time reduced any variation.

These results are limited to two longitudinal studies on Iris Murdoch and P.D. James. A longitudinal study of only two authors is a small sample, however, while the data highlighted depression from life events in Iris Murdoch and markers for Alzheimer's disease, they show promise, but they are not conclusive, even if Murdoch's depression and anxiety have been well documented through her prolific habit of writing about

herself throughout her life (Dooley & Nerlich, 2014; Martin & Rowe, 2010; Murdoch, 2015; Wilson, 2004), and there have been a number of scientific studies that verify it.

In the case of the findings of the Suicide Attacker study, when correlating the known times of depression in Iris Murdoch against the attackers one record was collected for each of the 25 attackers, so we cannot know whether individuals were depressed at the time they wrote their manifestos and final suicide notes. However, more research can be conducted using the Distress Analysis Interview Corpus (DAIC) (Gratch et al., 2014), which contains transcripts of clinical interviews designed to support the diagnosis of psychological distress such as anxiety, depression, and post-traumatic stress disorder.

11.6 Conclusions

This thesis contributed to the medieval literature authorship debate of Elizabethan playwrights and provided a new technique to separate works of disputed authorship. It introduced the idea that self or personality can be captured from a person's writing style using RPAS, a multi-disciplinary approach to identity. It found that sensory data can contribute to the early identification of Alzheimer's disease when used with critical slowing down to identify tipping points. It also found cognitive linguistic markers for depression and anxiety can be identified during troubled periods of a person's life from writing.

There were a number significant findings highlighted from the thesis:

1 – We determined that a multi-disciplinary approach to identity is an effective way to characterise a person's personality from writing.

2 – We were able to develop a new neuro-linguistic technique, RPAS, which is based on measures of a person's personality or self to create a signature of an individual's identity and find that it can separate people by their writing.

3 – We found that identity changes over time in an individual's writing due to natural aging (n=3).

4 – The use of highly concrete and imageability articles, pronouns, conjunctions, and prepositions were shown to relate to self, successfully predicting well-documented periods of depression and anxiety from writing (n = 2).

5 – We developed new techniques to visualise critical slowing down using RPAS and identify tipping points in individuals.

6 – We found it is possible to use sensory data from writing to detect linguistic markers of Alzheimer's disease (n=2).

7 – The use of modified variants of the 1-lag autocorrelation and Fischer-Pearson coefficient of skewness equations for critical slowing down can be used with RPAS to detect tipping points.

8 – Using RPAS and critical slowing down, we observed stylometric markers of Alzheimer's disease extending back 20 years before a formal medical diagnosis, which is earlier than results from Garrard *et al.* (2004), Le *et al.* (2011), and Hurst and Feng (2012).

9 – We determined that RPAS can separate the final manifestos and suicide notes of lone wolf suicide attackers from normal bloggers.

10 – We found that depression in an individual does not alter the RPAS classification of a person as a suicide attacker or 'normal' blogger.

11.7 Future Research

Many of the findings in this study have been exploratory, and therefore, this research could benefit from further testing to refine the RPAS algorithm. One area would be to retest the critical slowing down (CSD) phenomena that occur near tipping points by collecting data from one or more known terrorists across a period of their lives that includes writing just prior to their attack to see the point where any self-radicalisation might have occurred using the modified AR1 and G1 techniques. At this stage, our testing has only mimicked cognitive disorders, and testing real data from known terrorists would be beneficial. However, this is problematic in an unclassified arena due to attacker data being removed by agencies after a catastrophic event, often to stop copycat actions, or for legal reasons.

We also found that while chunking our data into file sizes of 4000 words provided good results in our contemporary author's dataset, it is possible that RPAS might perform better using a different file size, and this could be tested to optimise the algorithm further.

This thesis used a small dataset. It would benefit from using other modern data sources to further test the effectiveness of the RPAS algorithm to identify contemporary authors, including its ability to work against spoofing – the process where people write like another person or not like themselves to hide their identity. Internet Anonymity is becoming an increasingly sought-after concept (Faust *et al.*, 2017). While there are a number of Anonymous Social Networks (ASNs) that claim to provide anonymity and protect people's privacy (Day *et al.*, 2016), Brennan, Alfroz, and Greenstadt (2012) have developed techniques to conceal writing style, known as Adversarial Stylometry and avoid an author being recognised (Day *et al.*, 2016).

We have not tested the idea that somebody might be deliberately changing their language and their use of function words to evoke a different experience for the receiver (for propaganda or to hide their identity) and what this impact might be on the RPAS algorithm's effectiveness.

The RPAS algorithm would benefit from being placed within a mathematical modelling framework, where it could be automated to select input data and provide output charts. So, the research would benefit from mapping the elements of RPAS into a Bayesian Network model. Bayesian Belief Networks can be modelled to account for the influence of complex human behaviours and reduce the risk of decision making and uncertainty (Trucco *et al.*, 2008). Subjective logic, a powerful Bayesian reasoning model tool for conditional reasoning, is used in situations involving partial information and makes it possible to analyse Bayesian network models with uncertain probabilities (Jøsang, 2008). Subjective Logic is an extension of standard logic that uses continuous uncertainty and belief parameters instead of only discrete truth values and is suitable for handling uncertainty (Jøsang, 1997). Each of the RPAS element can be given a measure of effectiveness, both on each RPAS element generally, and on the particular piece of data being analysed. An overall confidence level could then be applied to the information to help in the classification of large datasets. It would make it easier to measure the distance between connections of different authors for social network analysis.

More work on authors with and without depression or cognitive disease could be conducted, such as using Patrick White (who did not have any reported cognitive decline) and Agatha Christie, British Prime Minister Harold Wilson, and US President Ronald Reagan (all of whom were reported to have cognitive decline in their later years). Outside of any real-world data on terrorists, this would strengthen the findings

on depression and the visualisation techniques for the Critical Slowing Down phenomena.

As an alternative to the sentiment tags used in LIWC for the research on suicide attackers, alternate datasets could be considered, such as the affective norms for English words (ANEW) database (Bradley & Lang, 1999), or Canada's National Research Council Sentiment and Emotion Lexicon (Mohammad & Turney, 2013).

Finally, the sensory processing of people over time could be extended to look at more types of diseases, such as Parkinson's disease where MCI is common prior to the development of a slightly different form of dementia than in AD (Lin & Wu, 2015). This could include off the cuff interviews and speeches from former world leaders, such as US President George H. W. Bush, Margaret Thatcher, and Canadian Prime Minister Pierre Elliot Trudeau, where there have been little to no edits by other parties.

References

Abbasi, A., & Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, *26*(2), 7.

Abdullah, L., & Rahim, A.F. (2016). A step-wise multiple linear regression analysis for identifying predictors of employees' intention to undertake further study. *Journal of Current Research in Science*, *4*(2), 21.

Ahmed, S., Haigh, A.M.F., de Jager, C.A., & Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*, *136*(12), 3727-3737.

Akbarzadeh, S. (2013). Investing in Mentoring and Educational Initiatives: The Limits of De-Radicalisation Programmes in Australia. *Journal of Muslim Minority Affairs*, *33*(4), 451-463.

Aljumily, R. (2015). Hierarchical and Non-Hierarchical Linear and Non-Linear Clustering Methods to "Shakespeare Authorship Question". *Social Sciences*, *4*(3), 758-799.

Ammar, D.Z.A., & Gohar Ayaz, A. (2016). Language Versus Thought, and Theory of Formation of Meanings. *Global Journal of Human-Social Science Research*, 15(12).

Amsel, B.D., Urbach, T.P., & Kutas, M. (2012). Perceptual and motor attribute ratings for 559 object concepts. *Behavior research methods*, *44*(4), 1028-1041.

Ancusa, V., Bogdan, R., & Caus, O. (2013). Enabling Knowledge Sharing in an Academic Environment: A Case Study. In Janiūnaitė, B., Pundziene, A., & Petraite, M. (Eds.). (2013, January). *Proceedings of the 14th European Conference on Knowledge Management: ECKM 2013*. Academic Conferences Limited.

Andersen, R.A., & Gnadt, J.W. (1989). Posterior parietal cortex. *The neurobiology of saccadic eye movements, ed. RH Wurtz & ME Goldberg. Elsevier*.

Andrei, A.L. (2014). Development and evaluation of Tagalog Linguistic Inquiry and Word Count (LIWC) dictionaries for negative and positive emotion. *The MITRE Corporation: Mclean, Virginia*.

Angus, D., Rintel, S., & Wiles, J. (2013). Making sense of big text: a visual-first approach for analysing text data using Leximancer and Discursis. *International Journal of Social Research Methodology*, *16*(3), 261-267.

Arefin, A.S., Vimieiro, R., Riveros, C., Craig, H., & Moscato, P. (2014). An information theoretic clustering approach for unveiling authorship affinities in Shakespearean era plays and poems. *PloS one*, *9*(10), e111445.

Argamon, S., Koppel, M., Fine, J., & Shimoni, A.R. (2003). Gender, genre, and writing style in formal written texts. *Text-the Hague then Amsterdam then Berlin-*, *23*(3), 321-346.

Argamon, S., Koppel, M., Pennebaker, J.W., & Schler, J. (2007). Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, *12*(9).

Argamon, S., Koppel, M., Pennebaker, J.W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2), 119-123.

Bakker, E. & DeGraaf, B. (2010). Lone wolves: How to prevent this phenomenon? *International Centre for Counter-Terrorism – The Hague*.

Bakker, E., & Roy, V. Z. J. D. (2015). Lone-Actor Terrorism: Definitional Workshop. *Countering Lone-Actor Terrorism Series*.

Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*.

Bandler, R., Grindler, J. (1979). 'Frogs into Princes'. *Moab, UT: Real People Press*, 1979.

Bargh, J.A., McKenna, K.Y., & Fitzsimons, G.M. (2002). Can you see the real me? Activation and expression of the "true self" on the Internet. *Journal of social issues*, *58*(1), 33-48.

Barnett, L., Lizier, J.T., Harré, M., Seth, A.K., & Bossomaier, T. (2013). Information flow in a kinetic Ising model peaks in the disordered phase. *Physical review letters*, *111*(17), 177203.

Barnfield, R. (1598). Lady Pecunia, Or, The Praise of Money: Also A Combat Betwixt Conscience and Covetousnesse; Together with The Complaint of Poetry for the Death of Liberality. In *Volume 1, Issue 7 of Illustrations of old English literature*. pp 1-49. Digitized 25 Oct 2012. Available at: https://books.google.com.au/books?id=0J1TAAAAcAAJ. Accessed on: 11 Nov 2015.

Barnfield, R. (1605). Lady Pecunia, Or, The Praise of Money: Also A Combat Betwixt Conscience and Covetousnesse; Together with The Complaint of Poetry for the Death of Liberality. In *Volume 1, Issue 4 of Illustrations of old English literature*. pp 1-38. Digitized 25 Oct 2012. Available at: https://books.google.com.au/books?id=y51TAAAAcAAJ. Accessed on: 11 Nov 2015.

Barnfield, R. (2008). Richard Barnfield. *The Project Gutenberg EBook of Encyclopaedia Britannica,* 11th edition, Volume 3, Part 1, Slice 3. Published 10 December, 2008. Page 415.

Barsalou, L.W., Simmons, W.K., Barbey, A.K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, *7*(2), 84-91.

Bartlett, M.S. (1950). Tests of significance in factor analysis. *British Journal of Mathematical and Statistical Psychology*, *3*(2), 77-85.

Bayley, J. (1998). Elegy for Iris. London: *Macmillan*.

Bayley, J. (1999). Iris: A Memoir of Iris Murdoch. New York: *Gerald Duckworth & Co Ltd*.

Bell, I. (2008). Rethinking Shakespeare's Dark Lady. *Schoenfeldt, A Companion*, 293-313.

Beale, E.M.L., Kendall, M.G., & Mann, D.W. (1967). The discarding of variables in multivariate analysis. *Biometrika*, *54*(3-4), 357-366.

Bechtel, W. (2002). Decomposing the mind-brain: A long-term pursuit. *Brain and Mind*, *3*(2), 229-242.

Beck, C.J., & Schoon, E. (2017). Terrorism and Social Movements; *Wiley-blackwell Companion to Social Movements*.

Bednarz, J.P. (2007) "Canonizing Shakespeare: The Passionate Pilgrim, England's Helicon and the Question of Authenticity," *Shakespeare Survey 60* (2007): 255-58,260,262.

Bednarz, J.P. (2012). The Mystery of 'The Phoenix and Turtle'. In *Shakespeare and the Truth of Love* (pp. 19-48). Palgrave Macmillan UK.

Benatti, F., & Tonra, J. (2015). English Bards and Unknown Reviewers: a Stylometric Analysis of Thomas Moore and the Christabel Review. *Breac: A Digital Journal of Irish Studies*.

Benjamin, V., Chung, W., Abbasi, A., Chuang, J., Larson, C.A., & Chen, H. (2014). Evaluating text visualization for authorship analysis. *Security Informatics*, *3*(1), 10.

Bentz, C., Kiela, D., Hill, F., & Buttery, P. (2014). Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. *Corpus Linguistics and Linguistic Theory*, *10*(2), 175-211.

Bermingham, A., Conway, M., McInerney, L., O'Hare, N., & Smeaton, A. F. (2009, July). Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. In *Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in* (pp. 231-236). IEEE.

Berisha, V., Wang, S., LaCross, A., & Liss, J. (2015). Tracking Discourse Complexity Preceding Alzheimer's Disease Diagnosis: A Case Study Comparing the Press Conferences of Presidents Ronald Reagan and George Herbert Walker Bush. *Journal of Alzheimer's Disease*.

Bhui, K., Everitt, B., & Jones, E. (2014). Might depression, psychosocial adversity, and limited social assets explain vulnerability to and resistance against violent radicalisation?. *PloS one*, *9*(9), e105918.

Bird, H., Ralph, M.A.L., Patterson, K., & Hodges, J.R. (2000). The Rise and Fall of Frequency and Imageability: Noun and Verb Production in Semantic Dementia. *Brain and Language*, 73, 17-49. https://doi.org/10.1006/brln.2000.2293

Bishara, A.J., & Hittner, J.B. (2012). Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological methods*, *17*(3), 399.

Blank, H., Kiebel, S.J., & von Kriegstein, K. (2015). How the human brain exchanges information across sensory modalities to recognize other people. *Human brain mapping*, *36*(1), 324-339.

Bloxham, A. (2011) Osama bin Laden's past video messages. *The Telegraph*. 19 May, 2011. Available at: http://www.telegraph.co.uk/news/worldnews/al-qaeda/8522949/Osama-bin-Ladens-past-video-messages.html Accessed on: 11 Sep 2014.

Bobadilla, L. (2014). Martyrdom redefined: Self-destructive killers and vulnerable narcissism. *Behavioral and brain sciences*, *37*(04), 364-365.

Bonin, P., Méot, A., Ferrand, L., & Bugaïska, A. (2015). Sensory experience ratings (SERs) for 1,659 French words: Relationships with other psycholinguistic variables and visual word recognition. *Behavior research methods*, *47*(3), 813-825.

Bourke, L., & Miller, N. (2017). London attack: Westminster attacker identified as British-born Khalid Masood. *The Sydney Morning Herald*, March 24, 2017. Available: http://www.smh.com.au/world/london-attack-westminster-attacker-identified-as-khalid-masood-20170323-gv592n.html. Accessed: 7 April, 2017.

Bordin. J. (2011) 'A crisis of trust and cultural incompatibility: A red team study of mutual perceptions of Afghan National Security Force personnel and US soldiers in understanding and mitigating the phenomena of ANSF committee fratricide-murders'. 12 May 2011. Available at: http://www.michaelyon-online.com/images/pdf/trust-incompatibility.pdf. Accessed on: 11 Feb 2012.

Bos, E.H., & De Jonge, P. (2014). "Critical slowing down in depression" is a great idea that still needs empirical proof. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(10), E878.

Bostwick, W.B., Meyer, I., Aranda, F., Russell, S., Hughes, T., Birkett, M., & Mustanski, B. (2014). Mental health and suicidality among racially/ethnically diverse sexual minority youths. *American journal of public health*, *104*(6), 1129-1136.

Bosward, M. (2017) 'Layers, Traces and Gaps: Collage, Found Footage and the Contested Past', *Presented at 'Animation and Memory', Radboud University, Nijmegen, 22-23 June*.

Boyd, R.L., & Pennebaker, J.W. (2015). Did Shakespeare write Double Falsehood? Identifying individuals by creating psychological signatures with text analysis. *Psychological science*, *26*(5), 570-582.

Braak, H., & Braak, E. (1995). Staging of Alzheimer's disease-related neurofibrillary changes. *Neurobiology of aging*, *16*(3), 271-278.

Bradley, M.M., & Lang, P.J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings (pp. 1-45). *Technical report C-1, the center for research in psychophysiology, University of Florida.*

Bragg, M. (2003). The Adventure of English. *Hodder and Stoughton*, London.

Brennan, M., Afroz, S., & Greenstadt, R. (2012). Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, *15*(3), 12.

Brewer, W.F. (1984). The story schema: Universal and culture-specific properties. *Center for the Study of Reading Technical Report; no. 322*.

Bringmann, L.F., Hamaker, E.L., Vigo, D.E., Aubert, A., Borsboom, D., & Tuerlinckx, F. (2017). Changing dynamics: Time-varying autoregressive models using generalized additive modeling. *Psychological methods*, *22*(3), 409.

Bristol, M.D. (1996). Big-Time Shakespeare. *Psychology Press.*

Brooke, T. (1922). The Marlowe Canon. *Publications of the Modern Language Association of America*, 367-417.

Brown, R., & Gilman, A. (1989). Politeness theory and Shakespeare's four major tragedies. *Language in society*, *18*(2), 159-212.

Brown, C., Snodgrass, T., Kemper, S.J., Herman, R., & Covington, M.A. (2008). Automatic Measurement of Propositional Idea Density from Part-of-Speech Tagging. *Behavior Research Methods*, 40, 540-545.https://doi.org/10.3758/BRM.40.2.540

Brunel, L., Carvalho, P.F., & Goldstone, R.L. (2015). It does belong together: cross-modal correspondences influence cross-modal integration during perceptual learning. *Frontiers in psychology*, *6*.

Bruster, D. (2013). Shakespearean spellings and handwriting in the additional passages printed in the 1602 Spanish Tragedy. *Notes and Queries*.

Brynielsson, J., Horndahl, A., Johansson, F., Kaati, L., Mårtenson, C., & Svenson, P. (2013). Harvesting and analysis of weak signals for detecting lone wolf terrorists. *Security Informatics*, *2*(1), 1-15.

Bucci, W. (1982). The vocalization of painful affect. *Journal of Communication disorders*, *15*(6), 415-440.

Bucci, W. (1984). Linking words and things: Basic processes and individual variation. *Cognition*, *17*(2), 137-153.

Bucci, W. (1997). Symptoms and symbols: A multiple code theory of somatization. *Psychoanalytic Inquiry*, *17*(2), 151-172.

Bucci, W. (2002). The referential Process, Consciousness, and the Sense of Self. *Psychoanalytic Inquiry: A Topical Journal for Mental Health Professionals*. Volume 22, Issue 5, 1 Nov 2002. PP. 766-793.

Bucci, W., and Freedman, N. (1981). The Language of Depression. *Bulletin of the Menninger Clinic*. 45: 34-358.

Bucci, W., Kabasakalian-McKay, R. (2004). Scoring Referential Activity: Instructions for Use with Transcripts of Spoken Texts. *Editor: Graham, E.A. Derner Institute Adelphi University, Garden City. N.Y. October 2004*. Pp. 24.

Bucci, W., & Maskit, B. (2004). Building a weighted dictionary for referential activity. In *Spring Symposium of the American Association for Artificial Intelligence in Palo Alto, CA, March*.

Bucci, W., & Maskit, B. (2006). A weighted referential activity dictionary. In *Computing attitude and affect in text: Theory and applications* (pp. 49-60). Springer Netherlands.

Bucci, W., Maskit, B., & Murphy, S. (2015). Connecting emotions and words: the referential process. *Phenomenology and the Cognitive Sciences*, 1-25.

Bucci, W., & Miller, N. E. (1993). Primary process analogue: The referential activity (RA) measure.

Bucholtz, M., & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, *7*(4-5), 585-614.

Buchta, C., Hornik, K., & Hahsler, M. (2008). Getting things in order: an introduction to the R package seriation. *Journal of Statistical Software*, *25*(3), 1-34.

Burnham, M. (1990). "Dark Lady and Fair man": The Love Triangle in Shakespeare's Sonnets and Ulysses. *Studies in the Novel*, 43-56. (43)

Burns, R.B., & Burns, R.A. (2012). Business Research Methods and Statistics Using SPSS. *SAGE Publications*.

Burrows, J., & Craig, H. (2012). Authors and characters. *English studies*, *93*(3), 292-309.

Byman, D. (2016). Omar Mateen, Lone-Wolf Terrorist. Slate. June 12, 2016. *Slate.* Accessed at: http://www.slate.com/articles/news_and_politics/foreigners/2016/06/lone_wolf_terrorists_like_omar_mateen_present_a_different_kind_of_threat.html. Accessed: 7 April, 2017.

Calvert, G.A., Campbell, R., & Brammer, M.J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current biology*, *10*(11), 649-657.

Capellan, J.A. (2015). Lone Wolf Terrorist or Deranged Shooter? A Study of Ideological Active Shooter Events in the US, 1970-2014. *Studies in Conflict & Terrorism*, *38*(6), 395-413.

Cechetto, D., & Weishaupt, N. (2016). The Cerebral Cortex in Neurodegenerative and Neuropsychiatric Disorders. *Academic Press*, 24 Feb. 2017.

Cerf-Ducastel, B., Van de Moortele, P.F., MacLeod, P., Le Bihan, D., & Faurion, A. (2001). Interaction of gustatory and lingual somatosensory perceptions at the cortical level in the human: a functional magnetic resonance imaging study. *Chemical Senses*, *26*(4), 371-383.

Chadefaux, T. (2014). Early warning signals for war in the news *Journal of Peace Research* January 2014 vol. 51 no. 1, 5-18

Champion, L.S. (1988). 'Answere to this perillous time': Ideological ambivalence in the raigne of king Edward III and the english Chronicle plays. *Journal of English Studies*. Vol. 69, 2, 117-129.

Chapman, S.J. (2017). Review of Discovering Statistics Using IBM SPSS Statistics. *Journal of Political Science Education*, Vol. 14, 1, 145-147.

Charmaz, K., & Mitchell, R.G. (1996). The myth of silent authorship: Self, substance, and style in ethnographic writing. *Symbolic Interaction*, *19*(4), 285-302.

Chaski, C.E. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International journal of digital evidence*, *4*(1), 1-13.

Chaski, C.E. (2001). Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, *8*, 1-65.

Chau, S.A., Chung, J., Herrmann, N., Eizenman, M., & Lanctôt, K.L. (2016). Apathy and Attentional Biases in Alzheimer's Disease. *Journal of Alzheimer's Disease*, (Preprint), 1-10.

Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., & Chau, M. (2004). Crime data mining: a general framework and some examples. *Computer*, *37*(4), 50-56.

Cheney, P. (2009). The Voice of the Author in 'The Phoenix and the Turtle:' Chaucer. *Shakespeare, Spenser." Perry and Watkins*, 103-25.

Cheng, N., Chandramouli, R., & Subbalakshmi, K.P. (2011). Author gender identification from text. *Digital Investigation*, *8*(1), 78-88.

Chiljan, K. (2012). Reclaiming The Passionate Pilgrim for Shakespeare. *Oxfordian*, 2012, Vol. 14, p74-81

Chung, C., Pennebaker, J. (2007) 'The Psychological Functions of Function Words'. In *K. Fiedler (Ed.) (2007). Social Communication (pp. 343-359) New York: Psychology Press.*

Chung, C.K., & Pennebaker, J.W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of research in personality*, *42*(1), 96-132.

Churchland, P.S. (2002). Self-representation in nervous systems. *Science*, *296*(5566), 308-310.

Coccia, M. (2008). Measuring scientific performance of public research units for strategic change. *Journal of Informetrics*, *2*(3), 183-194.

Cohen, S.J. (2016). Mapping the Minds of Suicide Bombers using Linguistic Methods: The Corpus of Palestinian Suicide Bombers' Farewell Letters (CoPSBFL). *Studies in Conflict & Terrorism*, *39*(7-8), 749-780.

Colbaugh, R., & Glass, K. (2012). Early warning analysis for social diffusion events. *Security Informatics*, *1*(1), 1-26.

Colaço, M., Mendonça, M., Farias, M., Henrique, P. (2010) 'OSS Developers Context-Specific Preferred Representational Systems: A Initial Neurolinguistic Text Analysis of the Apache Mailing List'. *Software Engineering Laboratory (LES)*, Federal University of Bahia, Salvador, Brazil.

Colaço Júnior, M., Mendonça, M., Farias, M., Henrique, P., & Corumba, D. (2012, November). A Neurolinguistic Method for Identifying OSS Developers' Context-Specific Preferred Representational Systems. In *ICSEA 2012, The Seventh International Conference on Software Engineering Advances* (pp. 112-121).

Coleridge, S.T. (1984). Biographia literaria, or, biographical sketches of my literary life and opinions (Vol. 7). *Princeton University Press*.

Coltheart, M. (1981). MRC Psycholinguistic database. In *Quarterly Journal of Experimental Psychology* 33A (1981):497-505.

Commonwealth of Australia (2002-14). General Intelligence. *Defence Intelligence Organisation*. Available at: http://www.defence.gov.au/dio/general-intelligence.shtml. Accessed 10 May 2015.

Commonwealth of Australia (2013). Defence White Paper 2013. *Department of Defence*, 2013.

Commonwealth of Australia (2014). Cyber 2020 Vision: DSTO cyber science and technology plan. *Defence Science and Technology Organisation, Department of Defence*, May 2014.

Commonwealth of Australia, (2015). Countering violent extremism. *Attorney-General's Department.* Available at: http://www.ag.gov.au/NationalSecurity/Counteringviolentextremism/Pages/default.aspx. Accessed 10 May 2015.

Connor, F.X. (2014). Shakespeare, poetic collaboration and The Passionate Pilgrim. pp119-130, in Holland, P. (Ed.). (2014). *Shakespeare Survey: Volume 67, Shakespeare's Collaborative Work* (Vol. 67). Cambridge University Press

Conway, M., & McInerney, L. (2008). Jihadi video and auto-radicalisation: Evidence from an exploratory YouTube study. In *Intelligence and Security Informatics* (pp. 108-118). Springer, Berlin, Heidelberg.

Cooper, A., (2016). Numan Haider had been radicalised before he stabbed police, inquest hears. *The Age*. March 11, 2016. Accessed at: http://www.theage.com.au/victoria/numan-haider-had-been-radicalised-before-he-stabbed-police-inquest-hears-20160311-gngxyt.html.Accessed: 7 April, 2017.

Cooper, H. (May 1, 2011). "Obama Announces Killing of Osama bin Laden". *The New York Times*. Retrieved 9 Sep 2014. Available at: http://thelede.blogs.nytimes.com/2011/05/01/bin-laden-dead-u-s-official-says/?_php=true&_type=blogs&_r=0

Cope, T.E., Rittman, T., Borchert, R.J., Jones, P.S., Vatansever, D., Allinson, K., ... & Rowe, J.B. (2017). Tau burden and the functional connectome in Alzheimer's disease and progressive supranuclear palsy. *Brain*.

Corballis, M.C. (2016). The evolution of language: sharing our mental lives. *Journal of Neurolinguistics, 43*, 120-132.

Cordy, J. (2017). The social media revolution: Political and security implications. *Draft CDS Report [064 CDSDG 17 E], NATO Parliamentary Assembly, Sub-Committee on Democratic Governance. Available online at www. nato-pa. int (last accessed September 7, 2017)*.

Corner, E., & Gill, P. (2015). A false dichotomy? Mental illness and lone-actor terrorism. *Law and human behavior*, *39*(1), 23.

Cox, J.M. (2010). DOMEX: The Birth of a New Intelligence Discipline. *Edited by Smith, S.A. In Intelligence. MIPB Military Intelligence Professional Bulletin* April-June 2010. PB 34-10-2.

Craig, H., & Kinney, A.F. (2009). *Shakespeare, computers, and the mystery of authorship*. Cambridge University Press.

Crosman, R. (1990). Making Love Out of Nothing At All: The Issue of Story in Shakespeare's Procreation Sonnets. *Shakespeare Quarterly*, 470-488.

Croy, I., Symmank, A., Schellong, J., Hummel, C., Gerber, J., Joraschky, P., & Hummel, T. (2014). Olfaction as a marker for depression in humans. *Journal of affective disorders*, 160, 80-86.

Dakos, V., Scheffer, M., van Nes, E. H., Brovkin, V., Petoukhov, V., & Held, H. (2008). Slowing down as an early warning signal for abrupt climate change. *Proceedings of the National Academy of Sciences*, *105*(38), 14308-14312.

Dakos, V., Van Nes, E. H., D'Odorico, P., & Scheffer, M. (2012). Robustness of variance and autocorrelation as indicators of critical slowing down. *Ecology*, *93*(2), 264-271.

Daly, F., Kramer, N., Mendlik, M., Pomerleau, E., & Ariemma, E. (2018). "Neither Gone nor Here": Coping with Personality Change and Loss of Identity in Neurologic Disease (FR415). *Journal of Pain and Symptom Management*, *55*(2), 601-602.

Dam, L. (2014). Mother-in-law, my, we know her!. *MenMand*, 887.

Damasio, A. (2003). Mental self: The person within. *Nature*, *423*(6937), 227-227.

Daugherty, L., & Press, C. (2011). The Assassination of Shakespeare's Patron: Investigating the Death of the Fifth Earl of Derby. *Brief Chronicles Vol. III (2011) ii*, 253.

Day, S., Brown, J., Thomas, Z., Bass, L., & Dozier, G. (2016, August). Adversarial Authorship, AuthorWebs, and Entropy-Based Evolutionary Clustering. In *Computer Communication and Networks (ICCCN), 2016 25th International Conference on* (pp. 1-6). IEEE.

Department of Defence (2016). Defence White Paper. *Commonwealth of Australia*, 2016.

Deshpande, G., Hu, X., Stilla, R., & Sathian, K. (2008). Effective connectivity during haptic perception: a study using Granger causality analysis of functional magnetic resonance imaging data. *Neuroimage*, *40*(4), 1807-1814.

De Vel, O., Anderson, A., Corney, M., and Mohay, G. (2001). 'Mining e-mail content for author identification forensics'. *ACM Sigmod Record*, 30(4), pp. 55-64, 2001.

Devington, D. (2007). The Poems by William Shakespeare. *Bantam Books*. New York.

Dickerson, B. C., & Eichenbaum, H. (2010). The episodic memory system: neurocircuitry and disorders. *Neuropsychopharmacology*, *35*(1), 86-104.

Dilts, R., Grindler, J., Bandler, R., DeLozier, J. (1980). 'NLP: The Study of the Structure of Subjective Experience, Volume 1'. *Meta Publications*, Capitola, CA.

Dooley, G., & Nerlich, G. (2014). Never Mind about the Bourgeoisie: The Correspondence between Iris Murdoch and Brian Medlin 1976-1995. Cambridge: *Cambridge Scholars Publishing*.

Drake, J. M., & Griffen, B.D. (2010). Early warning signals of extinction in deteriorating environments. *Nature*, *467*(7314), 456-459.

Drechsler, J., Bender, S., & Rässler, S. (2008). Comparing Fully and Partially Synthetic Datasets for Statistical Disclosure Control in the German IAB Establishment Panel. *Trans. Data Privacy*, *1*(3), 105-130.

Dreyer, F.R., & Pulvermüller, F. (2018). Abstract semantics in the motor system?–An event-related fMRI study on passive reading of semantic word categories carrying abstract emotional and mental meaning. *Cortex*, *100*, 52-70.

Driver, J., & Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. *Neuron*, *57*(1), 11-23.

Dugast, D. (1978). Sur quoi se fonde la notion d'étendue théorique du vocabulaire. *Français (Le) Moderne Paris*, *46*(1), 25-32.

Duncan-Jones, K. (1983). Was the 1609 Shake-speares Sonnets really unauthorized?. *Review of English Studies*, 151-171.

Durkheim, E. (1897) Le Suicide. 1897 France.

Earle, D., & Hurley, C.B. (2015). Advances in dendrogram seriation for application to visualization. *Journal of Computational and Graphical Statistics*, *24*(1), 1-25.

Efron, B., & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know?. *Biometrika*, *63*(3), 435-447.

Egnoto, M.J., & Griffin, D.J. (2016). Analyzing Language in Suicide Notes and Legacy Tokens. *Crisis*.

Elliott, W.E., & Valenza, R.J. (2010). Two tough nuts to crack: did Shakespeare write the 'Shakespeare'portions of Sir Thomas More and Edward III? Part I. *Literary and linguistic computing*, *25*(1), 67-83.

Elliot, W., & Valenza, R. (1991a). Was the Earl of Oxford the true Shakespeare. *Notes and Queries*, *38*(4), 501-506.

Elliott, W.E., & Valenza, R.J. (1991b). A Touchstone for the Bard. *Computers and the Humanities*, *25*(4), 199-209.

Ellis, D. (2000). Biography and Shakespeare: An outsider's view. *The Cambridge Quarterly*, *29*(4), 296-313.

Erne, L. (2013). Shakespeare and the book trade. *Cambridge University Press*. Pp. 56-86.

Esiri, M.M., & Wilcock, G.K. (1984). The olfactory bulbs in Alzheimer's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, *47*(1), 56-60.

Farey, P. (2014). Peter Farey's Marlowe Page. Available at: http://www2.prestel.co.uk/

Farrow, J.M. (1993). The Collected Works of Shakespeare. Available at: http://sydney.edu.au/engineering/it/~matty/Shakespeare/

Faust, C., Dozier, G., Xu, J., & King, M.C. (2017, November). Adversarial authorship, interactive evolutionary hill-climbing, and author CAAT-III. In *Computational Intelligence (SSCI), 2017 IEEE Symposium Series on* (pp. 1-8). IEEE.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874.

Ferguson, A., Spencer, E., Craig, H., & Colyvas, K. (2014). Propositional Idea Density in Women's Written Language over the Lifespan: Computerized Analysis. *Cortex*, 55, 107-121. https://doi.org/10.1016/j.cortex.2013.05.012

Fernandino, L., Binder, J.R., Desai, R.H., Pendl, S.L., Humphries, C.J., Gross, W.L., ... & Seidenberg, M.S. (2015). Concept Representation Reflects Multimodal Abstraction: A Framework for Embodied Semantics. *Cerebral Cortex*, *26*(5), 2018-2034.

Filote, A., Potrafke, N., & Ursprung, H. (2016). Suicide attacks and religious cleavages. *Public Choice*, *166*(1-2), 3-28.

Flekova, L., & Gurevych, I. (2013, September). Can we hide in the web? large scale simultaneous age and gender author profiling in social media. In *CLEF 2012 Labs and Workshop, Notebook Papers*.

Flottau, J. (2015). Opinion: Lessons To Learn From Germanwings Flight 9525. *Aviation Week & Space Technology*, 14 Apr 2015. Available at: http://aviationweek.com/commercial-aviation/opinion-lessons-learn-germanwings-flight-9525 . Accessed on 21 Apr 2015.

Fort, J.A. (1933). The Order and Chronology of Shakespeare's Sonnets. *Review of English Studies*, 19-23.

Foster, J., Bevis, M., & Businger, S. (2005). GPS Meteorology: Sliding-window analysis. *Journal of atmospheric and oceanic technology*, *22*(6), 687-695.

Frantzeskou, G., Stamatatos, E., Gritzalis, S., Chaski, C.E., & Howald, B.S. (2007). Identifying authorship by byte-level n-grams: The source code author profile (scap) method. *International Journal of Digital Evidence*, *6*(1), 1-18.

Fuentes, K.M. (2016). Cyberterrorism: The use of social networking to recruit Westerners an informational guide for law enforcement agencies (Doctoral dissertation, UTICA COLLEGE).

Gaonkara, B., Hovda, D., Martin, N., & Macyszyn, L. (2016, March). Deep learning in the small sample size setting: cascaded feed forward neural networks for medical image segmentation. In *SPIE Medical Imaging* (pp. 97852I-97852I). International Society for Optics and Photonics.

Garrard, P. (2009). Cognitive Archaeology: Uses, Methods, and Results. *Journal of Neurolinguistics*, 22, 250-265. https://doi.org/10.1016/j.jneuroling.2008.07.006

Garrard, P., Maloney, L.M., Hodges, J.R., & Patterson, K. (2005). The Effects of Very Early Alzheimer's disease on the Characteristics of Writing by A Renowned Author. *Brain*, 128, 250-260. https://doi.org/10.1093/brain/awh341

Gee, J.P., & Grosjean, F. (1984). Empirical evidence for narrative structure. *Cognitive Science*, *8*(1), 59-85.

Ghajar-Khosravi, S., Kwantes, P., Derbentseva, N., & Huey, L. (2016). Quantifying Salient Concepts Discussed in Social Media Content: A Case Study using Twitter Content Written by Radicalized Youth. *Journal of Terrorism Research*, *7*(2).

Gill, P., Horgan, J., & Deckert, P. (2014). Bombing alone: Tracing the motivations and antecedent behaviors of Lone-Actor terrorists. *Journal of Forensic Sciences, 59*(2), 425-435. doi:10.1111/1556-4029.12312

Gjelsvik, B., Lovric, D., & Williams, J.M.G. (2018). Embodied cognition and emotional disorders: Embodiment and abstraction in understanding depression. *Journal of Experimental Psychopathology*, *9*(3), pr-035714.

Gotelli, N.J., & Colwell, R.K. (2011). Estimating species richness. *Biological diversity: frontiers in measurement and assessment*, *12*, 39-54.

Gottfried, J.A., Deichmann, R., Winston, J.S., & Dolan, R.J. (2002). Functional heterogeneity in human olfactory cortex: an event-related functional magnetic resonance imaging study. *The Journal of Neuroscience*, *22*(24), 10819-10828.

Gratch, J., Artstein, R., Lucas, G.M., Stratou, G., Scherer, S., Nazarian, A., ... & Traum, D.R. (2014, May). The Distress Analysis Interview Corpus of human and computer interviews. In *LREC* (pp. 3123-3128).

Gray, H.D. (1920). The "Titus Andronicus" Problem. *Studies in Philology*, *17*(2), 126-131.

Gray, R.M. (2012). Science You Can Use: Neuroscience for Understanding and Expanding NLP Practices. *IASH Conference Presentation: September 16, 2012, Richard M. Gray, PHD School of Criminal Justice*.

Griffin, D.R. (2009). 'Osama Bin Laden: Dead Or Alive?'. *Interlink Books*.

Griffin, D.R. (2013). 'Osama bin Laden Responsible for the 9/11 Attacks? Where is the Evidence?' *Global Research*. Accessed: 11 Sep 2014. Available at: http://www.globalresearch.ca/osama-bin-laden-responsible-for-the-9-11-attacks-where-is-the-evidence/15892

Grindler, J., Bandler, R. (1976). 'The Structure of Magic II'. Palo Alto, Ca: *Science and Behavior Books*, 1976.

Gross, C. G. (1987). Neuroscience, early history of. *Journal of Enclyclopedia of Neuroscience,* pp. 843, 847.

Guell, X., Gabrieli, J.D., & Schmahmann, J.D. (2018). Embodied cognition and the cerebellum: Perspectives from the Dysmetria of Thought and the Universal Cerebellar Transform theories. *Cortex*, *100*, 140-148.

Guillot, A., Collet, C., Nguyen, V.A., Malouin, F., Richards, C., & Doyon, J. (2009). Brain activity during visual versus kinesthetic imagery: an fMRI study. *Human brain mapping*, *30*(7), 2157-2172.

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique*. Presses universitaires de France.

Guttal, V & Jayaprakash, C. (2008). Changing skewness: an early warning signal of regime shifts in ecosystems, *Ecology Letters*, 11, 2008, 450-460.

Hamm, M., & Spaaij, R. (2017). Broadcasting Intent: The Key to Preventing Lone Wolf Terrorism. In *The Age of Lone Wolf Terrorism* (pp. 91-121). New York: Columbia University Press. Retrieved from http://www.jstor.org.virtual.anu.edu.au/stable/10.7312/hamm18174.10

Hancock, A.B., Stutts, H.W., & Bass, A. (2014). Perceptions of Gender and Femininity Based on Language: Implications for Transgender Communication Therapy. *Language and Speech*, 0023830914549084.

Hancock, J.T., Landrigan, C., & Silver, C. (2007, April). Expressing emotion in text-based communication. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 929-932). ACM.

Harré, R. (1999a). 'Social Construction and Consciousness'. In *Investigating Phenomenal Consciousness: New Methodologies and Maps,* edited by Velmans, M. (ed.) (2000) Amsterdam: John Benjamins. (UK, USA), 2000, pp. 233-254.

Harré, R. (1999b). 'The rediscovery of the human mind: the discursive approach'. *Asian Journal of Social Psychology*, *2*(1), 43-62.

Hartshorne, J.K., & Germine, L.T. (2015). When Does Cognitive Functioning Peak? The Asynchronous Rise and Fall of Different Cognitive Abilities Across the Life Span. *Psychological science*, *26*(4), 433-443.

Hassan, S.U., Imran, M., Iftikhar, T., Safder, I., & Shabbir, M. (2017, November). Deep Stylometry and Lexical & Syntactic Features Based Author Attribution on PLoS Digital Repository. In *International Conference on Asian Digital Libraries* (pp. 119-127). Springer, Cham.

Heap, M. (1988a). 'Neurolinguistic Programming - An Interim Verdict'. In *Hypnosis: Current Clinical, Experimental and Forensic Practices*. London: Croom Helm, 1988, p. 268.

Heap, M. (1988b). 'Neuro-Linguistic Programming – A British Perspective'. In *Hypnos: Swedish Journal of Hypnosis in Psychotherapy and Psychosomatic Medicine*, Vol. 15, No.1, 1988, pp. 4-13.

Heap, M. (2008). 'The validity of some early claims of neuro-linguistic programming'. *Skeptical Intelligencer*, Vol. 11, pp. 6-13, 2008.

Heilman, K.M., Nadeau, S.E., & Beversdorf, D.O. (2003). Creative innovation: possible brain mechanisms. *Neurocase*, 9(5), 369-379.

Herring, S.C., Paolillo, J.C. (2006). 'Gender and genre variation in weblogs'. *Journal of Sociolinguistics*, 10(4): 439-459.

Hirsch, B.D., & Craig, H. (2014). "Mingled Yarn": The State of Computing in Shakespeare 2.0. *The Shakespearean International Yearbook*: 14, 3-35.

Hirst, G., & Wei Feng, V. (2012). Changes in style in authors with Alzheimer's disease. *English Studies*, *93*(3), 357-370.

Holdaway, J.C. (2014). Watching the Crew: Commercial Aircraft Operations and the Surveillance of Pilots before and after MH370. *The Arbutus Review*, *5*(1), 41-61.

Hoffmann, K. (2017, April). The psychology of the lone terrorist: Identification with the aggressor in individuals and in societies. In *International Forum of Psychoanalysis* (pp. 1-7). Routledge.

Hoffmann, S. (2004). Using the OED quotations database as a corpus–a linguistic appraisal. *ICAME journal*, *28*(4), 17-30.

Holm, H., Migne, M., Ahlse, E. (1994). Linguistic symptoms in dementia of Alzheimer type and their relation to linguistic symptoms of aphasia. *Logoped Phoniatr Vocol* 19, 99-106.

Horgan, J.G. (2017). Psychology of terrorism: Introduction to the special issue. *American Psychologist*, *72*(3), 199.

Horobin, S. (2010*). Studying the history of early English*. Palgrave Macmillan. London

Hota, S., Argamon, S., Koppel, M., Zigdon, I. (2006). 'Performing Gender: Automatic Stylistic Analysis of Shakespeare's Characters'. *Proc. Digital Humanities*, July 2006.

Hubin, E., Vanschoenwinkel, B., Broersen, K., De Deyn, P.P., Koedam, N., van Nuland, N.A., & Pauwels, K. (2016). Could ecosystem management provide a new framework for Alzheimer's disease?. *Alzheimer's & dementia*, *12*(1), 65-74.

Hunt, K.W. (1965). Grammatical Structures Written at Three Grade Levels. *NCTE Research Report No. 3.*

Hurley-Tesluk, M., McGuire, L., Schaie, K.W., & Willis, S.L. (1994, April). Change in word fluency over adulthood: A longitudinal linguistic cluster analysis. *Poster presented at the annual meeting of the American Psychological Society*, New York, NY, 1-17.

Illes, J. (1989). Neurolinguistic features of spontaneous language production dissociate three forms of neurodegenerative disease: Alzheimer's, Huntington's, and Parkinson's. *Brain and language*, *37*(4), 628-642.

Iqbal, F., Binsalleeh, H., Fung, B., & Debbabi, M. (2013). A unified data mining solution for authorship analysis in anonymous textual communications. *Information Sciences*, 231, 98-112.

Jackson, M.P. (2006). Shakespeare and the quarrel scene in arden of faversham. *Shakespeare Quarterly*, *57*(3), 249-293.

Jackson, M.P. (2005). Francis Meres and the Cultural Contexts of Shakespeare's Rival Poet Sonnets. *The Review of English Studies*, *56*(224), 224-246.

James, F.C., & Wamer, N.O. (1982). Relationships between temperate forest bird communities and vegetation structure. *Ecology*, 63(1), 159-171.

Jamrozik, A., McQuire, M., Cardillo, E.R., & Chatterjee, A. (2016). Metaphor: Bridging embodiment to abstraction. *Psychonomic bulletin & review*, 23(4), 1080-1089.

Jockers, M.L., & Witten, D.M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, *25*(2), 215-223.

Jones, B.W., & Chung, W. (2016, September). Topic modeling of small sequential documents: Proposed experiments for detecting terror attacks. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 310-312). IEEE.

Jøsang, A. (2008). Conditional reasoning with subjective logic. *Journal of Multiple-Valued Logic and Soft Computing*, *15*(1), 5-38.

Jøsang, A. (1997, December). Artificial reasoning with subjective logic. In *Proceedings of the second Australian workshop on commonsense reasoning on* (Vol. 48, p. 34). Perth.

Júnior, M.C., Mendonça, M., Farias, M., & Henrique, P. (2010, May). OSS developers context-specific Preferred Representational systems: A initial Neurolinguistic text analysis of the Apache mailing list. In *Mining Software Repositories (MSR), 2010 7th IEEE Working Conference on* (pp. 126-129). IEEE.

Junior, M.C., Farias, M.A.D.F., Maciel, I., Santos, P.H.D., & Mendonca, M. (2014, September). Triangulating Experiments in an Industrial Setting to Evaluate Preferred Representational Systems of Software Developers. In *Software Engineering (SBES), 2014 Brazilian Symposium on* (pp. 71-80). IEEE.

Juola, P. (2008). Authorship attribution. *Foundations and Trends® in Information Retrieval*, *1*(3), 233-334.

Juola, P., & Mikros, G.K. (2016). Cross-Linguistic Stylometric Features: A Preliminary Investigation.

Kågström, J., Kågström, E. & Karlsson, R. (2009). 'uClassify Gender Analyzer_v5'. Available at: http://www.uclassify.com/browse/uClassify/GenderAnalyzer_v5

Kahn, J.H., Tobin, R.M., Massey, A.E., & Anderson, J.A. (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American journal of psychology*, 263-286.

Kaiser, H.F. (1970). A second generation little jiffy. *Psychometrika*, *35*(4), 401-415.

Kambasković-Sawers, D. (2007). Three themes in one, which wondrous scope affords: Ambiguous Speaker and Storytelling in Shakespeare's Sonnets. *Criticism*, *49*(3), 285-305

Kambourakis, G. (2014). Anonymity and closely related terms in the cyberspace: An analysis by example. *Journal of information security and applications*, *19*(1), 2-17.

Kaminski, M.E. (2013). Real Masks and Real Name Policies: Applying Anti-Mask Case Law to Anonymous Online Speech. *Fordham Intellectual Property, Media & Entertainment Law Journal*, 23(815).

Karydis, M., & Tsirtsis, G. (1996). Ecological indices: a biometric approach for assessing eutrophication levels in the marine environment. *Science of the Total Environment*, *186*(3), 209-219.

Kavé, G., & Goral, M. (2016). Word retrieval in picture descriptions produced by individuals with Alzheimer's disease. *Journal of clinical and experimental neuropsychology*, *38*(9), 958-966.

Kayser, C., & Shams, L. (2015). Multisensory causal inference in the brain. *PLoS biology*, *13*(2), e1002075.

Kemper, S., Thompson, M., Marquis, J. (2001). Longitudinal change in language production: Effects of aging and dementia on grammatical complexity and propositional content. *Psychology and Aging* 16, 600.

Kernot, D. (2013). The Identification of unknown authors using cross-document co-referencing. Master of Philosophy thesis. *University of New South Wales.* Nov 2013.

Kernot, D. (2016). Can Three Pronouns Discriminate Identity in Writing. In *Sarker, R., Abbas, H., Dunstall, S., Kilby, P., Davis, R. Young, L. (eds) Data and Decision Sciences in Action: Proceedings of the Australian Society for Operations Research Conference 2016*, Springer, New York.

Kessler, B., Numberg, G., & Schütze, H. (1997, July). Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics* (pp. 32-38). Association for Computational Linguistics.

Khan, M.I., Sameem, B., Nikoui, V., & Dehpour, A.R. (2016). Is the war on terror induced-post traumatic stress disorder; the cause of suicide attack? An approach from psycho-cognitive and neurobiological perspective. *Advancements in Life Sciences*, 3(4), 109-111.

Kimura, D., & Tanaka-Ishii, K. (2014). Study on Constants of Natural Language Texts. *Information and Media Technologies*, *9*(4), 771-789.

Kintsch, W., & Keenan, J. (1973). Reading Rate and Retention as a Function of the Number of Propositions in the Base Structure of Sentences. *Cognitive Psychology*, 5, 257-274. https://doi.org/10.1016/0010-0285(73)90036-4

Kirby, A. (2007). The London bombers as "self-starters": A case study in indigenous radicalization and the emergence of autonomous cliques. *Studies in Conflict & Terrorism*, *30*(5), 415-428.

Klein, S. W. (1993). Speech lent by males: gender, identity, and the example of Stephen's Shakespeare. *James Joyce Quarterly*, *30*(3), 439-449.

Klimek, V., Stockmeier, C., Overholser, J., Meltzer, H. Y., Kalka, S., Dilley, G., & Ordway, G. A. (1997). Reduced levels of norepinephrine transporters in the locus coeruleus in major depression. *The Journal of neuroscience*, 17(21), 8451-8458.

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).

Koppel, M., Argamon, S., and Shimoni, A.R. (2002). 'Automatically categorizing written texts by author gender'. *Literary and Linguistic Computing*, 17(4), pp. 401-412.

Koppel, M., & Seidman, S. (2013). Automatically Identifying Pseudepigraphic Texts. In *EMNLP* (pp. 1449-1454).

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.

Korp, C. (2015). Shoemakers, Clowns, and Saints: The Narrative Afterlife of Thomas Deloney. Available at: http://escholarship.org/uc/item/8hk20311

Kosslyn, S. M. (1994). Image and brain: The resolution of the imagery debate. *MIT Press*, Cambridge, UK.

Kosslyn, S. M. (2005). Mental images and the brain. *Cognitive Neuropsychology*, *22*(3-4), 333-347.

Koyande, P. J., Deshmukh, D. H., & Naravadkar, S. D. (2016). T-Alert: Terrorist Alert System Using Data Mining Techniques. *International Journal of Engineering Science*, *5886*.

Kreeger, D. L. (1987). In re Shakespeare: The Authorship of Shakespeare on Trial: Preface. *Am. UL Rev.*, *37*, 609.

Krueger, A. B. (2007). What makes a terrorist. *Economics and the Roots of Terrorism, 6.*

Krsul, I., & Spafford, E. H. (1997). Authorship analysis: Identifying the author of a program. *Computers & Security*, *16*(3), 233-257.

Kubát, M., Matlach, V., & Cech, R. (2014). QUITA: Quantitative Index Text Analyzer. *Lüdenscheid: RAM-Verlag*.

Kubát, M., & Milička, J. (2013). Vocabulary richness measure in genres. *Journal of Quantitative Linguistics*, *20*(4), 339-349.

Lahneman, W. J. (2016). IC Data Mining in the Post-Snowden Era. *International Journal of Intelligence and CounterIntelligence*, *29*(4), 700-723.

Lai, Chao-Yue (2009). 'Author Gender Analysis'. Final project: from *I256 Applied Natural Language Processing, University of California*, Berkley, California, fall 2009. Accessed on: 11 Nov 2013. Available at: http://courses.ischool.berkeley.edu/i256/f09/Final%20Projects%20write-ups/LaiChaoyue_project_final.pdf

Lamb, A., Paul, M. J., & Dredze, M. (2013, June). Separating Fact from Fear: Tracking Flu Infections on Twitter. In *the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia,* 9-14 June 2013, 789 - 795).

Lancashire, I. (2010). Forgetful Muses: Reading the Author in the Text. *University of Toronto Press*. Toronto, Canada. https://doi.org/10.3138/9781442686328

Lancashire, I., & Hirst, G. (2009). Vocabulary Changes in Agatha Christie's Mysteries as an Indication of Dementia: A Case Study. In *19th Annual Rotman Research Institute Conference, Cognitive Aging: Research and Practice*, 8-10 March, 2010, Toronto, Canada. (pp. 8-10).

Lankford, A. (2014). Précis of The Myth of Martyrdom: What Really Drives Suicide Bombers, Rampage Shooters, and Other Self-Destructive Killers. *Behavioral and brain sciences*, *37*(04), 351-362.

Layton, R. M. (2016). An Examination of how Law Enforcement Assess Precursors of Radicalization by Observing the Social Media Content of Potential Terrorists: A Qualitative Case Study (*Doctoral dissertation, Northcentral University*).

Le, X. (2010). Longitudinal Detection of Dementia through Lexical and Syntactic Changes in Writing. *Master's Thesis. Toronto: Department of Computer Science, University of Toronto*. http://ftp.cs.toronto.edu/pub/gh/Le-MSc-2010.pdf

Le, X., Lancashire, I., Hirst, G., & Jokel, R. (2011). Longitudinal Detection of Dementia through Lexical and Syntactic Changes in Writing: A Case Study of Three British Novelists. *Literary and Linguistic Computing*, 26, 435-461. https://doi.org/10.1093/llc/fqr013

Leech, N. L., & Onwuegbuzie, A. J. (2007). An array of qualitative data analysis tools: A call for data analysis triangulation. *School psychology quarterly*, *22*(4), 557.

Lehnert, W. G. (1981). Plot units and narrative summarization. *Cognitive Science*, *5*(4), 293-331.

Leikin, M. (2016). What Do We Learn from Neurolinguistics?. In *The Palgrave Handbook of Economics and Language* (pp. 121-136). Palgrave Macmillan UK.

Lenard, D. B. (2014, January). Language and Gender in the 113th Congressional Speeches. In *Sociolinguistics Summer School 5*.

Lewis, J. A. (2005, March). The Internet and terrorism. In *Proceedings of the Annual Meeting (American Society of International Law)* (pp. 112-115). The American Society of International Law.

Li, M., Zeng, T., Liu, R., & Chen, L. (2013). Detecting tissue-specific early warning signals for complex diseases based on dynamical network biomarkers: study of type 2 diabetes by cross-tissue analysis. *Briefings in bioinformatics*, *15*(2), 229-243.

Lieberman, P. (2016). The evolution of language and thought. *Journal of Anthropological Sciences*, 94, 1-20.

Lin, C. H., & Wu, R. M. (2015). Biomarkers of cognitive decline in Parkinson's disease. *Parkinsonism & related disorders*, *21*(5), 431-443.

Linden, D. E., Prvulovic, D., Formisano, E., Völlinger, M., Zanella, F. E., Goebel, R., & Dierks, T. (1999). The functional neuroanatomy of target detection: an fMRI study of visual and auditory oddball tasks. *Cerebral Cortex*, *9*(8), 815-823.

Little, R. J. (1993). Statistical analysis of masked data. *Journal of Official statistics*, *9*(2), 407.

Litvinova, T., Seredin, P., Litvinova, O., & Zagorovskaya, O. (2016). Profiling a set of personality traits of text author: what our words reveal about us. *Research in Language*, *14*(4), 409-422.

Liiv, I. (2010). Seriation and matrix reordering methods: An historical overview. *Statistical analysis and data mining*, *3*(2), 70-91.

Liu, C.H.,Ma,X.,Wu,X.,Li,F.,Zhang,Y.,Zhou,F.C.,Wang,Y.J.,Tie,C.L.,Zhou,Z., Zhang, D.,Dong,J.,Yao,L.,Wang,C.Y., (2012). Resting-state abnormal baseline brain activity in unipolar and bipolar depression. *Neuroscience*. Letter 516, 202–206.

Lorés-Sanz, R. (2011). The construction of the author's voice in academic writing: The interplay of cultural and disciplinary factors. *Text & Talk-An Interdisciplinary Journal of Language, Discourse & Communication Studies*, *31*(2), 173-193.

Lynott, D., & Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behavior research methods*, *45*(2), 516-526.

Lynott, D., & Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Research Methods*, 2009, 41(2) 558-564.

Mahon, B.Z. (2015). What is embodied about cognition?. *Language, cognition and neuroscience*, *30*(4), 420-429.

Mair, D. (2016). # Westgate: A Case Study: How al-Shabaab used Twitter during an ongoing attack. *Studies in Conflict & Terrorism*.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics?. In *Computational Linguistics and Intelligent Text Processing* (pp. 171-189). Springer Berlin Heidelberg.

Mark, M. (2014). A Celebration of Women Writers. Available at: http://digital.libttrary.upenn.edu/women/cary/Mariam/Mariam.html Accessed 27 October 2014.

Martin, J. B., & Pechura, C. M. (Eds.). (1991). Mapping the Brain and Its Functions: *Integrating Enabling Technologies into Neuroscience Research* (Vol. 91, No. 8). National Academies Press.

Martin, P., & Rowe, A. (2010). Iris Murdoch: A Literary Life. *Palgrave Macmillan.* London. https://doi.org/10.1057/9780230282964

Mastin, L. (2011). The History of English: Middle English (c. 1100 – c. 1500) Available at: http://www.thehistoryofenglish.com/history_middle.html Accessed On: 15 Jun 2015.

Matsudaira, N. (1967). Some dynamical properties of the Ising ferromagnet. *Canadian Journal of Physics*, *45*(6), 2091-2111.

Matsuo, Y., & Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, *13*(01), 157-169.

Matthews, R. A., & Merriam, T. V. (1993). Neural computation in stylometry I: An application to the works of Shakespeare and Fletcher. *Literary and Linguistic Computing*, *8*(4), 203-209.

Matz, R. (2007). The world of Shakespeare's sonnets: an introduction. *McFarland*.

Maxim, J., Bryan, K., Thompson, I.M. (1994). Language of the Elderly: A Clinical Perspective, *Whurr Pub Ltd*., London.

May, S. (1972). Spenser's" Amyntas": Three Poems by Ferdinando Stanley, Lord Strange, Fifth Earl of Derby. *Modern Philology*, *70*(1), 49-52.

McGrath, C. (2003). 'Sexed Texts'. *The New York Times*, 10 August 2003. Available at: http://www.nytimes.com/2003/08/10/magazine/10WWLN.html Accessed on: 5 August 2011.

McIntosh, A. (2016). The Jackknife Estimation Method. *arXiv preprint arXiv:1606.00497*.

Meisel, C., Klaus, A., Kuehn, C., & Plenz, D. (2015). Critical Slowing Down Governs the Transition to Neuron Spiking. *PLoS computational biology*, *11*(2), e1004097-e1004097.

Meloy, J.R., & Gill, P. (2016). The lone-actor terrorist and the TRAP-18. *Journal of Threat Assessment and Management*, *3*(1), 37.

Mendenhall, T. C. (1887). The characteristic curves of composition. *Science*, *9*(214), 237-249.

Mendenhall, T. C. (1901). A mechanical solution of a literary problem. *Popular Science Monthly*, 60-2.

Menhinick, E. F. (1964). A comparison of some species-individuals diversity indices applied to samples of field insects. *Ecology*, 859-861.

Mercia, M.A, Johnson, M. (1984). 'Representational System Predicate Use and Convergence in Counciling: Gloria Revisited'. *Journal of Counseling Psychology*, Vol. 31, No. 2, p. 166.

Mergenthaler, E., Bucci, W. (1999). Linking verbal and non-verbal representations: Computer analysis of referential activity. *British Journal of Medical Psychology*; Sep, 1999, Vol 72 Part: 3 pp339-354.

Merriam, T. (1995). Possible light on a Kyd canon. *Notes and Queries*, *42*(3), 340-341.

Merriam, T. (1998). Heterogeneous authorship in early Shakespeare and the problem of Henry V. *Literary and Linguistic Computing*, *13*(1), 15-28.

Merriam, T. V., & Matthews, R. A. (1994). Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe. *Literary and Linguistic Computing*, *9*(1), 1-6.

Mesulam, M. M. (2003). Primary progressive aphasia—a language-based dementia. *New England Journal of Medicine*, 349(16), 1535-1542.

Metz, C. E. (1978, October). Basic principles of ROC analysis. In *Seminars in nuclear medicine* (Vol. 8, No. 4, pp. 283-298). Elsevier.

Metzl, J.M., & MacLeish, K.T. (2015). Mental illness, mass shootings, and the politics of American firearms. *American journal of public health*, *105*(2), 240-249.

Meyer, B., Ajchenbrenner, M., & Bowles, D. P. (2005). Sensory sensitivity, attachment experiences, and rejection responses among adults with borderline and avoidant features. *Journal of personality disorders*, *19*(6), 641-658.

Miller, G. A. (1995). The science of words. *Scientific American Library,* New York.

Mohammad, S. M., & Turney, P. D. (2013). NRC Emotion Lexicon. *NRC Technical Report*.

Mroszczyk, J. (2016). To die or to kill? An analysis of suicide attack lethality. *Terrorism and Political Violence*, 1-21.

Murdoch, I. (2016). Living on Paper: Letters from Iris Murdoch, 1934-1995. *A. Horner and A. Rowe (eds), Princeton, NJ: Princeton University Press.* https://doi.org/10.1515/9781400880300

Murphy, S., Maskit, B., & Bucci, W. (2015, June). Putting Feelings into Words: Cross-Linguistic Markers of the Referential Process. In *CLPsych@ HLT-NAACL* (pp. 80-88).

Nazif, A., Mohammed, N. I., Malakahmad, A., & Abualqumboz, M. S. (2016). Application of step wise regression analysis in predicting future particulate matter concentration episode. *Water, Air, & Soil Pollution*, *227*(4), 1-12.

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., & Woodard, D. (2017). Surveying Stylometry Techniques and Applications. *ACM Computing Surveys (CSUR)*, *50*(6), 86.

Nevalainen, T. (2006). *Introduction to Early Modern English*. Edinburgh University Press.

Newman, M.L., Pennebaker, J.W., Berry, D.S., & Richards, J.M. (2003). 'Lying words: Predicting deception from linguistic style'. *Personality and Social Psychology Bulletin*, 29, pp. 665–675 in Chung, C., Pennebaker, J. (2007) 'The Psychological Functions of Function Words', 2007 in K. Fiedler (Ed.) (2007). *Social Communication* (pp. 343-359) New York: Psychology Press.

Nicholas M, Obler LK, Albert ML, Helm-Estabrooks, N (1985). Empty speech in Alzheimer's disease and fluent aphasia. *J Speech Hear Res* **28**, 405-410.

Niedenthal, P. M., Barsalou, L. W., Winkielman, P., Krauth-Gruber, S., & Ric, F. (2005). Embodiment in attitudes, social perception, and emotion. *Personality and social psychology review*, *9*(3), 184-211.

Niedenthal, P. M. (2007). Embodying emotion. *science*, *316*(5827), 1002-1005.

Nilsson, M. (2017). Suicide–Dying for a higher purpose? A study of the motives of the Islamic female suicide bombers through Emile Durkheim's structural view on suicide.

Nonen, S., Kato, M., Takekita, Y., Wakeno, M., Sakai, S., Serretti, A., & Kinoshita, T. (2016). Polymorphism of rs3813034 in Serotonin Transporter Gene SLC6A4 Is Associated With the Selective Serotonin and Serotonin-Norepinephrine Reuptake Inhibitor Response in Depressive Disorder: Sequencing Analysis of SLC6A4. *Journal of clinical psychopharmacology*, 36(1), 27-31.

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain—a meta-analysis of imaging studies on the self. *Neuroimage*, *31*(1), 440-457.

Oosterwijk, S., Touroutoglou, A., Lindquist, K. A., Barrett, L.F., & Russell, J. A. (2015). The neuroscience of construction: what neuroimaging approaches can tell us about how the brain creates the mind. *The psychological construction of emotion. New York: Guilford*.

Pakhomov, S., Chacon, D., Wicklund, M., & Gundel, J. (2011). Computerized assessment of syntactic complexity in Alzheimer's disease: a case study of Iris Murdoch's writing. *Behavior Research Methods*, *43*(1), 136-144.

Pape, R.A. (2008). Dying to win: The strategic logic of suicide terrorism. In *The theory and practice of Islamic terrorism* (pp. 129-132). Palgrave Macmillan, New York.

Pennebaker, J. W. (2011). The secret life of pronouns. *New Scientist*, *211*(2828), 42-45.

Pennebaker, J.W., Booth, R.J., Boyd, R.L., & Francis, M.E. (2015). Linguistic Inquiry and Word Count: LIWC2015. *Pennebaker Conglomerates* (www.LIWC.net), Austin, TX.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015a). The development and psychometric properties of LIWC2015. *University of Texas at Austin.* Austin, TX. DOI:10.15781/T29G6Z

Pennebaker, J.W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R.J. (2007). The development and psychometric properties of LIWC2007. www.LIWC.Net.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*, 2001.

Pennebaker, J. W., & Lay, T. C. (2002). Language use and personality during crises: Analyses of Mayor Rudolph Giuliani's press conferences. *Journal of Research in Personality*, 36(3), 271-282.

Pennebaker, J.W., Mehl, M.R., and Niederhoffer, K.G. (2003). Psychological Aspects of Natural Language Use**:** Our Words, Our Selves. *In The Psychology of Word Use. Annual Review of Psychology*, 2003. Issue 54 pp547-577.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.

Pilla, E. (2012). Sonnet 19, in *Eds. Baker, W., & Womack, K. The Facts On File Companion to Shakespeare, 5-Volume Set*. Infobase Pub. 282-284.

Poiret, R., & Liu, H. (2017). Mastering the measurement of text's frequency structure: an investigation on Lambda's reliability. *Glottometrics 37*, 82.

Popescu, V (2014). The many Faces of Love: A (Re) Reading of Shakespeare's Sonnets Four Centuries later. *Communication, Context, Interdisciplinarity*, Volume 3 2014. 281-290.

Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). Visual dominance: An information-processing account of its origins and significance. *Psychological review*, *83*(2), 157.

Post, F. (1996). Verbal creativity, depression and alcoholism. An investigation of one hundred American and British writers. *The British Journal of Psychiatry*, *168*(5), 545-555.

Pontone, G. M., Mari, Z., Perepezko, K., Weiss, H. D., & Bassett, S. S. (2016). Personality and reported quality of life in Parkinson's disease. *International journal of geriatric psychiatry*.

Powers, D. M. (1998, January). Applications and explanations of Zipf's law. In *Proceedings of the joint conferences on new methods in language processing and computational natural language learning* (pp. 151-160). *Association for Computational Linguistics*.

Preotiuc-Pietro, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., ... & Ungar, L. (2015). The Role of Personality, Age and Gender in Tweeting about Mental Illnesses. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, NAACL*.

Priest, A. (2013). You and I listening to me: towards an understanding of the significance of personal pronoun usage in psychotherapy *(Doctoral dissertation, Middlesex University and the Metanoia Institute*).

Putney, R. (1941). Venus and Adonis: Amour with Humor. *Philological Quarterly*, 20, 533.

Quarmley, M., Moberg, P. J., Mechanic-Hamilton, D., Kabadi, S., Arnold, S. E., Wolk, D. A., & Roalf, D. R. (2017). Odor Identification Screening Improves Diagnostic Classification in Incipient Alzheimer's Disease. *Journal of Alzheimer's Disease*, *55*(4), 1497-1507.

Raghunath, R., Balamuthusamy, S., Nguyen, P., Vallurupalli, A., Afolabi, O., & Bireddy, S. (2016). Clinical Predictors of recurrent stenosis and the need for re-intervention in the cephalic arch in patients with brachio-cephalic AV fistulas. Available at: https://digitalcommons.hsc.unt.edu/rad/RAD16/GeneralMedicine/4/. Accessed: 8 October, 2018.

Rajan, K. B., Wilson, R. S., Weuve, J., Barnes, L. L., & Evans, D. A. (2015). Cognitive Impairment 18 Years before Clinical Diagnosis of Alzheimer Disease Dementia. *Neurology*, 85, 898-904. https://doi.org/10.1212/WNL.0000000000001774

Raju, N. V., Kumar, V. V., & Rao, O. S. (2016). Author Based Rank Vector Coordinates (ARVC) Model for Authorship Attribution. *International Journal of Image, Graphics & Signal Processing*, *8*(5), 68.

Ralph, M.A.L., Jefferies, E., Patterson, K., & Rogers, T.T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42.

Ralston, N., Benny-Morrison, A., & Olding, R. (2015). Parramatta shooting: Gunman identified as Farhad Khalil Mohammad Jabar. *The Sydney Morning Herald*, October 4, 2015. Available: http://www.smh.com.au/nsw/parramatta-shooting-gunman-identified-as-farhad-jabar-khalil-mohammad-20151003-gk0jze.html. Accessed: 7 April, 2017.

Ramakrishna, K. (2014). Countering the self-radicalised lone wolf: a new paradigm?. (RSIS Commentaries, No. 019). *RSIS Commentaries. Singapore: Nanyang Technological University*.

Ramirez, J. (2016). Suicide: Across the Life Span. *Nursing Clinics of North America*, 51(2), 275-286.

Ramirez-Esparza, N., Chung, C. K., Kacewicz, E., & Pennebaker, J. W. (2008). The Psychology of Word Use in Depression Forums in English and in Spanish: Texting Two Text Analytic Approaches. In *ICWSM*.

Rangel, F., & Rosso, P. (2013). Use of language and author profiling: Identification of gender and age. *Natural Language Processing and Cognitive Science*, 177.

Ray, S. (2016). How social media is changing the way people commit crimes and police fight them. *USApp–American Politics and Policy Blog*.

Reardon, S. (2015). Science seeks roots of terror: psychological studies raise prospect of intervention in the radicalization process. *Nature*, *517*(7535), 420-422.

Reid, L. A. (2012). "Certaine Amorous Sonnets, Betweene Venus and Adonis": fictive acts of writing in The Passionate Pilgrime of 1612. *Études Épistémè. Revue de littérature et de civilisation (XVIe–XVIIIe siècles).*

Reuters (2009). "Q+A: Noordin Mohammad Top and Islamic militancy in Indonesia". *Reuters*. 2009-09-17. Accessed 11 Sep 2014. Available at: http://www.reuters.com/article/2009/09/17/us-indonesia-militants-qanda-sb-idUSTRE58G2OR20090917

Richards, I. A. (1958). The Sense of Poetry: Shakespeare's" The Phoenix and the Turtle". *Daedalus*, *87*(3), 86-94.

Rickards, B. (2014). Sonnets of Shakespeare by William Shakespeare. *Encyclopedia of Literature*. Salem Press, January, 2014.

Rigby, P., Hassan, A. (2007). 'What Can OSS Mailing Lists Tell Us? A Preliminary Psychometric Text Analysis of the Apache Developer Mailing List'. *Proceedings of the Fourth MSR*, 2007.

Robert, P., Bremond, F., & David, R. (2016). Depression, apathy and Alzheimer's disease: new perspectives. *Neurobiology of Aging*, 39, S29-S30.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, *12*(1), 77.

Rodriguez, J. D., Perez, A., & Lozano, J. A. (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, *32*(3), 569-575.

Rosenstein, M., Foltz, P. W., DeLisi, L. E., & Elvevåg, B. (2015). Language as a biomarker in those at high-risk for psychosis. *Schizophrenia research*, *165*(2), 249-250.

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of official Statistics*, *9*(2), 461-468.

Rubin, D. C., Deffler, S. A., Ogle, C. M., Dowell, N. M., Graesser, A. C., & Beckham, J. C. (2016). Participant, rater, and computer measures of coherence in posttraumatic stress disorder. *Journal of Abnormal Psychology*, 125, 11e25. http://dx.doi.org/10.1037/abn0000126.

Rubinstein, I. S., Lee, R. D., & Schwartz, P. M. (2008). Data mining and Internet profiling: Emerging regulatory and technological approaches. *The University of Chicago Law Review*, *75*(1), 261-285.

Rude, S.S., Gortner, E.M., and Pennebaker, J.W. (2004). Language Use of Depressed and Depression-Vulnerable College Students. *Cognition & Emotion*. 18: 1121-1133.

Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31, 351–365.

Rudman, J. (2012). The state of non-Traditional authorship attribution studies—2012: Some problems and solutions. *English Studies*, *93*(3), 259-274.

Rudman, J. (2016). Non-Traditional Authorship Attribution Studies of William Shakespeare's Canon: Some Caveats. *Journal of Early Modern Studies*, *5*, 307-328.

Sabat, S. R., & Harré, R. (1992). The construction and deconstruction of self in Alzheimer's disease. *Ageing and Society*, 12(04), 443-461.

Saif, H., Dickinson, T., Kastler, L., Fernandez, M., & Alani, H. (2017, May). A Semantic Graph-Based Approach for Radicalisation Detection on Social Media. In *European Semantic Web Conference* (pp. 571-587). Springer, Cham.

Saikal, A. (2015). What should we call Islamic State: DAISH or IS? *The Canberra Times*, January 18, 2015. http://www.canberratimes.com.au/comment/what-should-we-call-islamic-state-daish-or-is-20150118-12sii7.html Accessed 10 May 2015.

Sarma, K. M. (2017). Risk assessment and the prevention of radicalization from nonviolence into terrorism. *American Psychologist*, *72*(3), 278.

Sauer, M. M. (2008). The Facts on File Companion to British Poetry Before 1600. *Infobase Publishing*.

Sawyer, R. (2017). The Twenty-First Century: "Trauma, Drama, Conspiracy". In *Marlowe and Shakespeare* (pp. 307-341). Palgrave Macmillan, New York.

Scheffer, M. (2010). Complex systems: foreseeing tipping points. *Nature*, 467(7314), 411-412.

Schiermeier, Q. (2015). Attempts to predict terrorist attacks hit limits. *Nature*, *517*, 419-420.

Schuurman, B., Bakker, E., Gill, P., & Bouhana, N. (2018). Lone actor terrorist attack planning and preparation: A Data-driven analysis. *Journal of Forensic Sciences, 63*(4), 1191-1200. doi:10.1111/1556-4029.13676

Segarra, S., Eisen, M., Egan, G., & Ribeiro, A. (2017). Stylometric analysis of early modern period English plays. *Digital Scholarship in the Humanities, 1-8-2017, 1-18*.

Seidman, S. (2013). Authorship verification using the impostors method. In *CLEF 2013 Evaluation Labs and Workshop-Online Working Notes*.

Sellars, R. W. (1916). Critical realism. *Russell & Russell*, New York.

Sellars, R. W. (1959). II.—Sensations as guides to perceiving. *Mind*, *68*(269), 2-15.

Sellars, R. W. (1961). Referential transcendence. *Philosophy and Phenomenological Research*, 1-15.

Shakespeare, W., & Melchiori, G. (1998). King Edward III. *Cambridge University Press*.

Shannon, C. E. (1948). A mathematical theory of communication, *Bell System Technical Journal*, **27**: 379–423, 623–656.

Sharpley, C. (1984). 'Predicate Matching in NLP: a Review of Research on the Preferred Representational System'. *Journal of Counseling Psychology*, 31, pp. 238-248.

Sharpley, C. (1987). 'Research Findings on Neurolinguistic Programming: Non-supportive Data or Untestable Theory?'. *Journal of Counseling Psychology*, 34, pp. 103-107.

Shears, R. (2009). Terrorist mastermind Noordin Mohammed Top shot dead after a 17-hour siege. *Daily Mail Australia*. Accessed: 11 Sep 2014. Available at: http://www.dailymail.co.uk/news/article-1205164/Terrorist-mastermind-Noordin-Mohammed-Top-shot-dead-17-hour-seige.html

Sheehan, I. S. (2014). Are suicide terrorists suicidal? A critical assessment of the evidence. *Innovations in clinical neuroscience*, *11*(9-10), 81-92.

Shepherd, W.R. (1926). The Historical Atlas. Available http://www.lib.utexas.edu/maps/historical/history_middle_east.html. Accessed 10 May, 2015.

Shimojo, S., & Shams, L. (2001). Sensory modalities are not separate modalities: plasticity and interactions. *Current opinion in neurobiology*, *11*(4), 505-509.

Simon, H. J., & Wiese, H. (Eds.). (2002). *Pronouns Grammar and Representation* (Vol. 52). John Benjamins Publishing.

Simonto, D. K. (1989). Shakespeare's Sonnets: A Case of and for Single–Case Historiometry. *Journal of Personality*, *57*(3), 695-721

Simpson, E. (2014). How can we stop more attacks? *Political Science Publications.* Paper 54. http://ir.lib.uwo.ca/politicalsciencepub/54

Singhal, A., Buckley, C., & Mitra, M. (1996, August). Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 21-29). ACM.

Skillicorn, D. B., Alsadhan, N., Billingsley, R., & Williams, M. A. (2017). Social Robot Modelling of Human Affective State. *arXiv preprint arXiv:1705.00786*.

Slater, D. (2012). Early warning signals of tipping-points in blog posts. *The MITRE corporation, Virginia, USA*.

Smith, C. H. (2016). The linguistics of terror: A content analysis of suicide notes and martyr manifestos. (*Doctoral dissertation, The University of Alabama*).

Smith, S.R., Chenery, H.J., Murdoch, B.E. (1989). Semantic abilities in dementia of the Alzheimer type. II. Grammatical semantics. *Brain Lang,* 36, 533-542.

Smith, S. (2012). 'Paper tabled by the Minister for Defence on Afghanistan'. *Department of Defence*, Australian Government, 31 October, 2012. Available at: http://www.minister.defence.gov.au/2012/10/31/minister-for-defence-paper-tabled-by-the-minister-for-defence-on-afghanistan/ Accessed on: 2 November 2012.

Snowden, J., Griffiths, H., & Neary, D. (1994). Semantic dementia: Autobiographical contribution to preservation of meaning. *Cognitive Neuropsychology*, *11*(3), 265-288.

Snowdon, D. A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., & Markesbery, W. R. (1996). Linguistic Ability in Early Life and Cognitive Function and Alzheimer's Disease in Late Life: Findings from the Nun Study. *JAMA*, 275, 528-532. https://doi.org/10.1001/jama.1996.03530310034029

Spaaij, R. (2012). Understanding lone-wolf terrorism: Global patterns, motivations, and prevention. New York*: Springer*.

Speckhard, A., & Akhmedova, K. (2006). Black widows: The Chechen female suicide terrorists. *Female suicide bombers: Dying for equality*, *84*(1), 63-80.

Speed, L. J., & Majid, A. (2016). Dutch modality exclusivity norms: Simulating perceptual modality in space. *Behavior Research Methods*, *49*(6), 2204-2218.

St Clair, R. N. (2017). The Biological Rationale for Revising Communication Theory: Mirror Neurons, Epigenetics, Brain Functions, and Lexicon-based Semantics. *Intercultural Communication Studies*, *26*(1).

Stahel, W., Maechler, M. (2011). 'Jitter' (Add Noise) to Numbers. *R Documentation* (1995 – 2011) available at: http://stat.ethz.ch/R-manual/R-devel/library/base/html/jitter.html. Accessed: 2 August 2016.

START: National Consortium for the Study of Terrorism and Responses to Terrorism (START). (2016). Global Terrorism Database [Data file]. Retrieved from https://www.start.umd.edu/gtd

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

Stanczyk, U., Cyran, K.A. (2007). 'Machine learning approach to authorship attribution of literary texts'. *International Journal of Applied Mathematics and Informatics*, Issue 4, Volume 1, 2007, pp. 151-158.

Stapleton, M. L. (1993). "My False Eyes": The Dark Lady and Self-Knowledge. *Studies in Philology*, 213-230. Page 224.

Stevens, T. (2009). Regulating the 'Dark web': How a two-fold approach can tackle peer-to-peer radicalisation. *The RUSI Journal*, *154*(2), 28-33.

Stirman, S.W., and Pennebaker, J.W. (2001). Word Use in the Poetry of Suicidal and Non-Suicidal Poets. *Psychosomatic Medicine*. 63: 517-522.

Stockwell, P., Colomb, R. M., Smith, A. E., & Wiles, J. (2009). Use of an automatic content analysis tool: A technique for seeing both local and global scope. *International Journal of Human-Computer Studies*, *67*(5), 424-436.

Stottlemyre, S. (2014). The Effect of Country-Level Income on Domestic Terrorism: A Worldwide Analysis of the Difference between Lone-Wolf and Group Affiliated Domestic Terrorism. (Doctoral dissertation, Georgetown University).

Stritmatter, R. (2004). Law Case in Verse: Venus and Adonis and the Authorship Question, A. *Tenn. L. Rev.*, *72*, 171.

Sun, Y., Hunt, S., & Sah, P. (2015). Norepinephrine and corticotropin-releasing hormone: partners in the neural circuits that underpin stress and anxiety. *Neuron*, 87(3), 468-470.

Swaim, C. (2017). Establishing a stylometric baseline for micro-attributions of Shakespeare's apocrypha with 'On a day, alack the day'. In *Proceedings: 13th Annual Symposium on Graduate Research and Scholary Projects. Wichita, KS: Wichita State University*, p87.

Szot, P. (2016). Elevated Cerebrospinal Fluid Norepinephrine in the Elderly can Link Depression and A Reduced Glymphatic System as Risk Factors for Alzheimer's disease. *Journal of Aging Science*.

Tan, H. Z. (2000). Perceptual user interfaces: haptic interfaces. *Communications of the ACM*, *43*(3), 40-41.

Tanaka-Ishii, K., & Aihara, S. (2015). Computational Constancy Measures of Texts—Yule's K and Rényi's Entropy. *Computational Linguistics*.

Tardif, E., Doudin, P. A., & Meylan, N. (2015). Neuromyths Among Teachers and Student Teachers. *Mind, Brain, and Education*, *9*(1), 50-59.

Tarlinskaja, M. (2006). Shakespeare Among Others in Edward III and Sir Thomas More: From Meter to Authorship. Seattle, Washington.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24-54.

Tearle, M., Taylor, K., & Demuth, H. (2008). An algorithm for automated authorship attribution using neural networks. *Literary and linguistic computing*, *23*(4), 425-442.

The Sydney Morning Herald (2013). Samantha Lewthwaite: shy schoolgirl to terrorism suspect. *The Sydney Morning Herald.* Published 26 September 2013. Available at: http://www.smh.com.au/world/samantha-lewthwaite-shy-schoolgirl-to-terrorism-suspect-20130925-2udsn.html . Accessed 8 October 2018.

Thisted, R., & Efron, B. (1987). Did Shakespeare write a newly-discovered poem? *Biometrika*, *74*(3), 445-455.

Thomas, M. W. (2000). Eschewing credit: Heywood, Shakespeare, and plagiarism before copyright. *New Literary History*, *31*(2), 277-293.

Thomson, G. (2015). Magic in Practice: Introducing Medical NLP: the art and science of language in healing and health. *Hammersmith Books Limited*.

Tillman, R., & Louwerse, M. (2018). Estimating Emotions Through Language Statistics and Embodied Cognition. *Journal of psycholinguistic research*, *47*(1), 159-167.

Tosey, P., & Mathison, J. (2010). Exploring inner landscapes through psychophenomenology: The contribution of neuro-linguistic programming to innovations in researching first person experience. *Qualitative Research in Organizations and Management: An International Journal*, *5*(1), 63-82.

Toutanova, K., & Manning, C. D. (2000, October). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13 (pp. 63-70). Association for Computational Linguistics, Hong Kong, 3-6 October, 2000*.

Trefois, C., Antony, P. M., Goncalves, J., Skupin, A., & Balling, R. (2015). Critical transitions in chronic disease: transferring concepts from ecology to systems medicine. *Current opinion in biotechnology*, *34*, 48-55.

Trucco, P., Cagno, E., Ruggeri, F., & Grande, O. (2008). A Bayesian Belief Network modelling of organisational factors in risk analysis: A case study in maritime transportation. *Reliability Engineering & System Safety*, *93*(6), 845-856.

Tucker, E., Gelineau, K. (2016). Sydney siege gunman Man Haron Monis attracted attention of FBI in 2009. *The Sydney Morning Herald*, May 19, 2016. Available: http://www.smh.com.au/nsw/sydney-siege-gunman-man-haron-monis-attracted-attention-of-fbi-in-2009-20160518-goygo5.html. Accessed: 7 April, 2017.

Tuldava, J. (2004). The development of statistical stylistics (a survey). *Journal of Quantitative Linguistics*, *11*(1-2), 141-151.

Turner, A., & Greene, E. (1977). The Construction and Use of a Propositional Text Base (Technical Report 63). Boulder, CO: *University of Colorado, Institute for the Study of Intellectual Behavio*r.

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical Richness in perspective. *Computers and the Humanities*, 32(5), 323-352.

van Dantzig, S., Cowell, R. A., Zeelenberg, R., & Pecher, D. (2011). A sharp image or a sharp knife: Norms for the modality-exclusivity of 774 concept-property items. *Behavior research methods*, *43*(1), 145-154.

van de Leemput, I. A., Wichers, M., Cramer, A. O., Borsboom, D., Tuerlinckx, F., Kuppens, P., ... & Derom, C. (2014). Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences*, *111*(1), 87-92.

Van Gijsel, S., Speelman, D., & Geeraerts, D. (2005). A variationist, corpus linguistic analysis of lexical richness. *Proceedings of Corpus Linguistics 2005*.

Van Velzen, M., & Garrard, P. (2008). From Hindsight to Insight-Retrospective Analysis of Language Written by a Renowned Alzheimer's Patient. *Interdisciplinary Science Reviews*, 33, 278-286. https://doi.org/10.1179/174327908X392852

van Velzen, M. H., Nanetti, L., & de Deyn, P. P. (2014). Data modelling in corpus linguistics: How low may we go? *Cortex*, *55*, 192-201.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language testing*, *17*(1), 65-83.

Vickers, B. (2014). "The Two Authors of Edward III". *Shakespeare Survey. Ed. Peter Holland. 1st ed. Vol. 67. Cambridge: Cambridge University Press, 2014. pp. 102-118.* Shakespeare Survey Online. Web. 05 March 2015. http://dx.doi.org.virtual.anu.edu.au/10.1017/SSO9781107775572.008

Vickers, B. (2011). Shakespeare and authorship studies in the twenty-first century. *Shakespeare Quarterly*, *62*(1), 106-142.

Vickers, B. (2007). Shakespeare, 'A Lover's Complaint', and John Davies of Hereford. *Cambridge University Press*.

Voorhees, E. M. (1998). Using WordNet for text retrieval. *WordNet: an electronic lexical database*, 285-303.

Walther, B.A., Morand, S. (1998). Comparative performance of species richness estimation methods. *Cambridge University Press.*

Weintraub W. (1981). Verbal Behavior: Adaptation and Psychopathology. New York: *Springer*.

Weir, G.R., Dos Santos, E., Cartwright, B., & Frank, R. (2016, June). Positing the problem: enhancing classification of extremist web content through textual analysis. In *Cybercrime and Computer Forensic (ICCCF), IEEE International Conference on (pp. 1-3). IEEE.*

Wesson, D.W., Wilson, D.A., & Nixon, R.A. (2010). Should olfactory dysfunction be used as a biomarker of Alzheimer's disease?. *Expert review of neurotherapeutics*, *10*(5), 633-635.

Whorf, B.L. (1997). The relation of habitual thought and behavior to language. In *Sociolinguistics* (pp. 443-463). Macmillan Education UK.

Wichers, M., Borsboom, D., Tuerlinckx, F., Kuppens, P., Viechtbauer, W., van de Leemput, I. A., ... & Scheffer, M. (2014). Reply to Bos and De Jonge: Between-subject data do provide first empirical support for critical slowing down in depression. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(10), E879.

Wicklund, A.H., Johnson, N., & Weintraub, S. (2004). Preservation of reasoning in primary progressive aphasia: further differentiation from Alzheimer's disease and the behavioral presentation of frontotemporal dementia. *Journal of Clinical and Experimental Neuropsychology*, 26(3), 347-355.

Williams, H. (2005). Cassell's Chronology of World History. London: *Weidenfeld & Nicolson*. pp. 233–238.

Wilson, A. (2014). Tortured past that taught PD James the darkness inside the human heart: As the crime genius (and scourge of the BBC) dies, a friend's heartfelt tribute. *The Daily Mail.* 28 November, 2014. Available at: http://www.dailymail.co.uk/news/article-2852411/Tortured-past-taught-PD-James-darkness-inside-human-heart-crime-genius-scourge-BBC-dies-friend-s-heartfelt-tribute.html#ixzz4XEshrwLi. Accessed on: 08 October, 2018.

Wilson, A.N. (2004). Iris Murdoch as I Knew Her. New York: *Random House*.

Wilson, M. (1987). MRC Psycholinguistic Database: Machine Usable Dictionary. Version 2.00. April 29, 1987, Informatics Division. *Science and Engineering Research Council.*

Wilson, R.R. (1988). Shakespearean Narrative: The Rape of Lucrece Reconsidered. *Studies in English literature, 1500-1900*, *28*(1), 39-59.

Wimmer, G., & Altmann, G. (1999). Review article: On vocabulary richness. *Journal of Quantitative Linguistics*, *6*(1), 1-9.

Winkielman, P., Ziembowicz, M., & Nowak, A. (2015). The coherent and fluent mind: How unified consciousness is constructed from cross-modal inputs via integrated processing experiences. *Frontiers in Psychology*, *6*, 83.

Witkowski, T. (2010). Thirty-five years of research on Neuro-Linguistic Programming. NLP research data base. State of the art or pseudoscientific decoration? *Polish Psychological Bulletin*, *41*(2), 58-66.

Wittgenstein, L. (1922). 'Tractatus Logico-Philosophicus'. *Kegan Paul, Trench, Trubner & Co., Ltd.* New York, 1922.

Woodward, M.R., Dwyer, M.G., Amrutkar, C.V., Zivadinov, R., & Szigeti, K. (2015). Olfactory Identification Deficit as a Predictor of White Matter Tract Integrity in Alzheimer's Disease. *The American Journal of Geriatric Psychiatry*, *23*(3), S103-S104.

Woudhuysen, H. R. (1996). Sir Philip Sidney and the circulation of manuscripts, 1558-1640. *Oxford University Press.*

Wright, S., Denney, D., Pinkerton, A., Jansen, V., & Bryden, J. (2016). Resurgent Insurgents: Quantitative Research Into Jihadists Who Get Suspended but Return on Twitter. *Journal of Terrorism Research*, *7*(2).

Wright, W.R., & Chin, D.N. (2014, July). Personality profiling from text: introducing part-of-speech N-grams. In *International Conference on User Modeling, Adaptation, and Personalization (pp. 243-253). Springer International Publishing*.

Yan, M., Zhang, W.H., Wang, H., & Wong, K.M. (2017, November). The Dynamics of Bimodular Continuous Attractor Neural Networks with Moving Stimuli. In *International Conference on Neural Information Processing* (pp. 648-657). Springer, Cham.

Yang, A., Peng, C.K., & Goldberger, A.L. (2017, February). The Marlowe-Shakespeare Authorship Debate: Approaching an Old Problem with New Methods. Web.

Yang, C.C., & Ng, T.D. (2007, May). Terrorism and crime related weblog social network: Link, content analysis and information visualization. In *Intelligence and Security Informatics, 2007 IEEE (pp. 55-58). IEEE.*

Ye, J., Janardan, R., & Li, Q. (2004). Two-dimensional linear discriminant analysis. In *Advances in neural information processing systems* (pp. 1569-1576).

Yim, O., & Ramdeen, K.T. (2015). Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. *Quant. Methods. Psycholo*gy, *11*, 8-21.

Yoo, S.S., Freeman, D.K., McCarthy III, J.J., & Jolesz, F.A. (2003). Neural substrates of tactile imagery: a functional MRI study. *Neuroreport*, *14*(4), 581-585.

Zabelina, D.L., O'Leary, D., Pornpattananangkul, N., Nusslock, R., & Beeman, M. (2015). Creativity and sensory gating indexed by the P50: Selective versus leaky sensory gating in divergent thinkers and creative achievers. *Neuropsychologia*, *69*, 77-84.

Zou, Y.M., Lu, D., Liu, L.P., Zhang, H.H., & Zhou, Y.Y. (2016). Olfactory dysfunction in Alzheimer's disease. *Neuropsychiatric disease and treatment*, *12*, 869.

Zhao, Y., & Zobel, J. (2007, January). Searching with style: Authorship attribution in classic literature. In *Proceedings of the thirtieth Australasian conference on Computer science-Volume 62 (pp. 59-68). Australian Computer Society, Inc.*

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.

Appendix A

Table 33: Referential Activity Power data

WORD	TYPE	RA
A	A	0.29857
ABOUT	P	0.32286
AFTER	P	0.32786
ALL	PNOUN	0.42786
ALTHOUGH	C	0.30429
AN	A	0.29429
AND	C	0.31857
ANY	PNOUN	0.30571
ANYBODY	PNOUN	0.48
AS	PNOUN	0.27286
AT	PNOUN	0.29929
BAR	P	0.82929
BECAUSE	C	0.31429
BEFORE	P	0.36857
BEST	PNOUN	0.43286
BOTH	PNOUN	0.44286
BUT	PNOUN	0.30929
BY	P	0.31214
CAUSE	C	0.40643
CROSS	P	0.725
DOWN	P	0.57
EACH	PNOUN	0.42
ELSE	C	0.30714
EVERY	PNOUN	0.37929
EXTRA	P	0.42786
FAILING	P	0.51143
FEW	PNOUN	0.45857
FOR	P	0.32929
FORE	P	0.45714
FORTH	P	0.47
FROM	P	0.31143
GIN	P	0.85286
HE	PNOUN	0.60786
HER	PNOUN	0.63071
HIM	PNOUN	0.56357
HIMSELF	PNOUN	0.43
HIS	PNOUN	0.46929

WORD	TYPE	RA
ONCE	C	0.455
ONE	PNOUN	0.57929
ONLY	P	0.33786
OR	P	0.33714
OTHER	PNOUN	0.38571
OUR	PNOUN	0.4
OUT	P	0.41357
OVER	P	0.43857
OWN	PNOUN	0.46714
PAST	P	0.52786
POST	P	0.75071
ROUND	P	0.71214
SAME	PNOUN	0.41
SAVE	P	0.485
SELF	PNOUN	0.66071
SHE	PNOUN	0.62857
SIN	P	0.51
SINCE	P	0.365
SO	C	0.30143
SOME	PNOUN	0.42
SOUTH	P	0.58786
STILL	C	0.49929
SUCH	PNOUN	0.33357
THAN	C	0.28643
THAT	PNOUN	0.32643
THE	A	0.31857
THEIR	PNOUN	0.36571
THEM	PNOUN	0.50857
THEN	C	0.28143
THESE	PNOUN	0.37643
THEY	PNOUN	0.44857
THIS	PNOUN	0.37071
THOSE	PNOUN	0.35786
THROUGH	P	0.42429
THWART	P	0.47429
TILL	C	0.52071
TO	P	0.29929

I	PNOUN	0.67714
IF	C	0.30857
IN	P	0.43714
INN	P	0.83571
INTO	P	0.40357
IT	PNOUN	0.38286
ITS	PNOUN	0.31286
LESS	P	0.39429
LIKE	P	0.45571
ME	PNOUN	0.67214
MINE	PNOUN	0.69571
MORE	P	0.39143
MY	PNOUN	0.39357
NEAR	P	0.53214
NEXT	P	0.46286
NONE	PNOUN	0.50929
NOR	C	0.30929
NOW	C	0.41786
OF	P	0.28357
OFF	P	0.43357
ON	P	0.37857

UNDER	P	0.53214
UP	P	0.565
UPON	P	0.46071
US	PNOUN	0.58929
VICE	P	0.58
WE	PNOUN	0.47071
WHAT	PNOUN	0.38714
WHEN	PNOUN	0.32143
WHERE	C	0.365
WHICH	PNOUN	0.33714
WHILE	C	0.36143
WHO	PNOUN	0.35357
WHOM	PNOUN	0.38857
WHY	C	0.37
WITH	P	0.39643
WITHOUT	C	0.40643
WITHOUT	P	0.40643
YET	C	0.34714
YONDER	PNOUN	0.49214
YOU	PNOUN	0.55
YOUR	PNOUN	0.38929

List of function words and their RA Power values, where Type A = Article, C = Conjunctive, P = Preposition, and PNOUN = Pronoun.

Table 34: Sensory Adjectives data

Word	Modality	Exclusivity
abrasive	visual	0.469512
abrasive	haptic	0.433761
absorbent	auditory	0.401216
absorbent	visual	0.479784
aching	haptic	0.647287
aching	haptic	0.640741
acidic	gustatory	0.494737
acidic	gustatory	0.471178
acrid	olfactory	0.434641
acrid	gustatory	0.474619
adhesive	haptic	0.448666
adhesive	haptic	0.48072
alcoholic	olfactory	0.437198
alcoholic	gustatory	0.422535
amber	visual	0.636364
amber	visual	0.749035
aromatic	olfactory	0.510582
aromatic	olfactory	0.62585
astringent	gustatory	0.462427
astringent	olfactory	0.402632
azure	visual	0.616935
azure	visual	0.689498
babbling	auditory	0.505102
babbling	auditory	0.440514
balmy	haptic	0.356688
balmy	haptic	0.252492
banging	visual	0.385776
banging	auditory	0.456422
barbecued	olfactory	0.157601
barbecued	olfactory	0.180068
barking	auditory	0.536524
barking	auditory	0.530864
beautiful	visual	0.516971
beautiful	auditory	0.699634
beeping	auditory	0.615337
beeping	auditory	0.623907
beery	olfactory	0.497076
beery	olfactory	0.216354
beige	visual	0.785441
beige	visual	0.915493
big	visual	0.748062
big	visual	0.392276

Word	Modality	Exclusivity
icy	visual	0.35412
immense	haptic	0.640601
immense	visual	0.89823
insipid	gustatory	0.379187
insipid	auditory	0.604839
itchy	haptic	0.62776
itchy	haptic	0.644599
jagged	visual	0.534435
jagged	visual	0.441176
jammy	gustatory	0.259601
jammy	gustatory	0.289428
jingling	auditory	0.480638
jingling	auditory	0.5
juicy	gustatory	0.301887
juicy	gustatory	0.230769
khaki	visual	0.644068
khaki	visual	0.659722
large	visual	0.484694
large	visual	0.441237
laughing	visual	0.515957
laughing	auditory	0.522788
leathery	haptic	0.329389
leathery	haptic	0.377404
lemony	gustatory	0.439331
lemony	olfactory	0.655629
light	olfactory	0.717131
light	haptic	0.472362
lilting	auditory	0.639391
lilting	visual	0.529727
lithe	visual	0.727273
lithe	visual	0.416235
long	visual	0.665625
long	auditory	0.668342
loose	visual	0.488491
loose	visual	0.483333
loud	auditory	0.570225
loud	visual	0.497268
low	auditory	0.7
low	visual	0.569405
lukewarm	haptic	0.518625
lukewarm	gustatory	0.348115
lumpy	visual	0.305147

bitter	gustatory	0.530026
bitter	haptic	0.483745
black	visual	0.512195
black	visual	0.877551
black and white	visual	0.879464
black and white	visual	0.73617
bland	gustatory	0.499106
bland	visual	0.769811
blaring	auditory	0.613181
blaring	auditory	0.594901
bleating	auditory	0.516129
bleating	auditory	0.524476
bleeping	auditory	0.593865
bleeping	auditory	0.531507
bloody	visual	0.456057
bloody	visual	0.378486
blotchy	visual	0.665595
blotchy	visual	0.651389
blue	visual	0.932692
blue	visual	0.910638
blunt	haptic	0.541916
blunt	haptic	0.444604
boiling	auditory	0.308772
boiling	visual	0.185185
booming	auditory	0.4447
booming	auditory	0.412664
bouncy	haptic	0.477064
bouncy	visual	0.436428
branching	visual	0.57622
branching	visual	0.606918
braying	auditory	0.412616
braying	auditory	0.493151
breakable	visual	0.417582
breakable	visual	0.412346
breezy	auditory	0.242126
breezy	haptic	0.371901
bright	visual	0.981481
bright	visual	0.707719
brilliant	visual	0.729167
brilliant	auditory	0.840164
briny	gustatory	0.353391
briny	visual	0.221631
bristly	haptic	0.38785
bristly	haptic	0.47773
lumpy	haptic	0.514234
lush	visual	0.433628
lush	gustatory	0.265976
meaty	visual	0.453165
meaty	gustatory	0.299669
mellow	auditory	0.730612
mellow	gustatory	0.44697
melted	visual	0.427208
melted	visual	0.230279
metallic	gustatory	0.312989
metallic	visual	0.534574
mild	gustatory	0.470862
mild	visual	0.419825
miniature	visual	0.578947
miniature	visual	0.452214
minty	gustatory	0.422907
minty	olfactory	0.502907
moaning	auditory	0.586806
moaning	auditory	0.491139
moist	gustatory	0.314841
moist	haptic	0.318264
motionless	haptic	0.391421
motionless	visual	0.528395
mottled	visual	0.418457
mottled	visual	0.394191
mouldy	visual	0.39798
mouldy	visual	0.38322
muddy	visual	0.341991
muddy	visual	0.356137
murky	visual	0.443031
murky	visual	0.461728
murmuring	auditory	0.608187
murmuring	auditory	0.515464
mushroomy	visual	0.508621
mushroomy	gustatory	0.393574
mushy	haptic	0.226714
mushy	haptic	0.351145
musty	olfactory	0.469208
musty	olfactory	0.291221
narrow	visual	0.660436
narrow	visual	0.511393
noisy	auditory	0.545685
noisy	auditory	0.759717
nutty	gustatory	0.219055

brittle	haptic	0.391192
brittle	haptic	0.452309
broad	visual	0.543354
broad	visual	0.723164
broken	visual	0.328814
broken	visual	0.415842
bronze	visual	0.601329
bronze	visual	0.558912
brown	visual	0.736641
brown	visual	0.793358
bubbling	visual	0.237288
bubbling	auditory	0.3
bulky	visual	0.474074
bulky	visual	0.484485
bumpy	haptic	0.402
bumpy	haptic	0.450602
burning	visual	0.37747
burning	olfactory	0.217929
burnt	haptic	0.320704
burnt	gustatory	0.228614
bursting	auditory	0.404124
bursting	visual	0.377593
buttery	haptic	0.454343
buttery	gustatory	0.281365
buzzing	auditory	0.528455
buzzing	auditory	0.539823
caramelised	gustatory	0.2443
caramelised	gustatory	0.241796
charred	visual	0.271868
charred	visual	0.200918
cheesy	gustatory	0.289515
cheesy	olfactory	0.479254
chequered	visual	0.794574
chequered	visual	0.745923
chewy	gustatory	0.334677
chewy	gustatory	0.314629
chilly	haptic	0.46832
chilly	haptic	0.428571
chiming	auditory	0.559229
chiming	auditory	0.60479
chirping	auditory	0.665517
chirping	auditory	0.585366
chocolatey	gustatory	0.345576
chocolatey	visual	0.507808
nutty	gustatory	0.412037
odorous	olfactory	0.523929
odorous	olfactory	0.510753
oily	gustatory	0.272866
oily	visual	0.389706
oniony	visual	0.452872
oniony	gustatory	0.367311
open	visual	0.407336
open	visual	0.335185
orange	visual	0.91866
orange	visual	0.395299
oval	visual	0.508816
oval	visual	0.622951
painful	auditory	0.570621
painful	haptic	0.557229
pale	visual	0.837302
pale	visual	0.962617
patterned	visual	0.713311
patterned	visual	0.662539
peachy	gustatory	0.476427
peachy	visual	0.525
peppery	gustatory	0.437367
peppery	gustatory	0.375472
perfumed	olfactory	0.432432
perfumed	olfactory	0.501597
petite	visual	0.620178
petite	visual	0.626546
pink	visual	0.669903
pink	visual	0.951111
plain	visual	0.617555
plain	gustatory	0.414894
plastic	visual	0.355967
plastic	haptic	0.281955
polished	visual	0.417051
polished	auditory	0.682482
popping	auditory	0.414942
popping	auditory	0.412998
portly	visual	0.329377
portly	visual	0.59542
prickly	haptic	0.48642
prickly	haptic	0.563177
puffy	haptic	0.474299
puffy	visual	0.542029
pulsing	visual	0.702929

chubby	visual	0.517073
chubby	visual	0.490566
circular	visual	0.513661
circular	visual	0.433486
citrusy	olfactory	0.48125
citrusy	gustatory	0.4375
clammy	haptic	0.548387
clammy	visual	0.288381
clamorous	auditory	0.36253
clamorous	auditory	0.444444
clanging	auditory	0.537468
clanging	auditory	0.484988
clean	olfactory	0.3186
clean	visual	0.460526
clear	visual	0.677193
clear	visual	0.752727
clicking	auditory	0.455847
clicking	auditory	0.486216
cloudy	visual	0.536634
cloudy	visual	0.644951
cloying	haptic	0.504582
cloying	gustatory	0.461538
coarse	haptic	0.418478
coarse	haptic	0.485531
coconutty	gustatory	0.434053
coconutty	olfactory	0.501377
cold	gustatory	0.356589
cold	haptic	0.42539
colorful	visual	0.945833
colorful	visual	0.922414
colossal	visual	0.734483
colossal	visual	0.536649
compact	visual	0.558824
compact	haptic	0.482972
conical	visual	0.432304
conical	visual	0.590747
contoured	visual	0.414948
contoured	visual	0.473239
cooing	auditory	0.47027
cooing	auditory	0.573431
cool	haptic	0.440299
cool	haptic	0.382766
crackling	auditory	0.336576
crackling	auditory	0.321839
pulsing	haptic	0.467192
pungent	olfactory	0.537572
pungent	olfactory	0.459854
purple	visual	0.773723
purple	visual	0.915584
purring	auditory	0.483029
purring	auditory	0.485175
quiet	auditory	0.553672
quiet	auditory	0.514986
radiant	visual	0.677656
radiant	visual	0.413408
rancid	gustatory	0.371041
rancid	olfactory	0.455338
raspy	auditory	0.891667
raspy	haptic	0.385027
rectangular	visual	0.512887
rectangular	visual	0.788491
red	visual	0.845833
red	visual	0.567568
reddish	visual	0.804878
reddish	visual	0.943299
resounding	auditory	0.614089
resounding	auditory	0.585714
reverberating	auditory	0.731449
reverberating	auditory	0.388672
rhythmic	auditory	0.413462
rhythmic	auditory	0.562682
ripe	gustatory	0.342803
ripe	gustatory	0.279035
rippled	visual	0.382423
rippled	visual	0.436782
roaring	auditory	0.526923
roaring	auditory	0.540404
roasted	visual	0.20205
roasted	olfactory	0.2411
rotten	visual	0.267564
rotten	olfactory	0.321859
rough	visual	0.616352
rough	haptic	0.444976
round	visual	0.482436
round	visual	0.719665
rubbery	gustatory	0.325188
rubbery	haptic	0.377454
rumbling	auditory	0.49589

craggy	visual	0.522818
craggy	visual	0.571429
crashing	visual	0.459135
crashing	visual	0.302655
creaking	auditory	0.491443
creaking	auditory	0.529399
creamy	haptic	0.396588
creamy	gustatory	0.371595
creased	visual	0.425882
creased	visual	0.495775
crimson	visual	0.580716
crimson	visual	0.855183
crinkled	visual	0.450739
crinkled	visual	0.326568
crisp	olfactory	0.278676
crisp	gustatory	0.197514
crooked	visual	0.557971
crooked	visual	0.55914
crowded	visual	0.34965
crowded	visual	0.347584
crunching	auditory	0.252087
crunching	auditory	0.425968
crying	auditory	0.419831
crying	visual	0.417391
curly	visual	0.363758
curly	visual	0.525381
curved	visual	0.492537
curved	visual	0.521303
cute	visual	0.428899
cute	visual	0.442516
damp	haptic	0.370098
damp	haptic	0.353783
dank	visual	0.336538
dank	visual	0.346864
dappled	visual	0.699422
dappled	visual	0.736909
dark	visual	0.405405
dark	visual	0.753623
dazzling	visual	0.911628
dazzling	visual	0.505319
dead	visual	0.383795
dead	auditory	0.629758
deafening	auditory	0.571429
deafening	auditory	0.722222

rumbling	auditory	0.458904
rustling	auditory	0.373303
rustling	auditory	0.431034
rusty	visual	0.395876
rusty	visual	0.364238
salty	gustatory	0.465854
salty	gustatory	0.418803
savory	gustatory	0.332779
savory	gustatory	0.370927
scaly	haptic	0.366013
scaly	haptic	0.465218
scented	olfactory	0.662252
scented	olfactory	0.44802
scratchy	auditory	0.502591
scratchy	haptic	0.515625
scrawny	visual	0.614907
scrawny	visual	0.539945
screaming	auditory	0.517966
screaming	auditory	0.58011
screeching	auditory	0.637462
screeching	auditory	0.531328
shadowy	visual	0.787149
shadowy	visual	0.776423
shaggy	visual	0.5
shaggy	haptic	0.455422
shallow	auditory	0.544025
shallow	visual	0.435768
sharp	gustatory	0.415217
sharp	haptic	0.509822
sheer	visual	0.690377
sheer	visual	0.504505
shimmering	visual	0.533569
shimmering	visual	0.677419
shiny	visual	0.651163
shiny	visual	0.64557
short	visual	0.553977
short	visual	0.558333
shrieking	auditory	0.542284
shrieking	auditory	0.528967
shrill	auditory	0.462766
shrill	auditory	0.829787
silky	haptic	0.516971
silky	haptic	0.516854
silver	visual	0.741935

deep	visual	0.930036
deep	visual	0.471883
delicious	gustatory	0.354724
delicious	olfactory	0.657439
dim	visual	0.254613
dim	visual	0.936937
dirty	visual	0.32037
dirty	visual	0.373757
downy	visual	0.417625
downy	visual	0.462484
drab	visual	0.587333
drab	visual	0.504801
dry	gustatory	0.364641
dry	visual	0.410138
dull	visual	0.779592
dull	auditory	0.556391
dusty	visual	0.314928
dusty	visual	0.441805
earthy	visual	0.568643
earthy	gustatory	0.283019
echoing	auditory	0.649231
echoing	auditory	0.789343
eggy	olfactory	0.523179
eggy	olfactory	0.288889
elastic	haptic	0.496124
elastic	haptic	0.407407
elegant	visual	0.574286
elegant	visual	0.57377
empty	visual	0.378906
empty	visual	0.32906
enormous	visual	0.476551
enormous	visual	0.540682
faint	visual	0.290323
faint	olfactory	0.5
falling	visual	0.735409
falling	visual	0.204412
fat	visual	0.503817
fat	visual	0.27707
fatty	gustatory	0.343612
fatty	visual	0.346392
fetid	olfactory	0.374724
fetid	olfactory	0.324503
feverish	visual	0.828704
feverish	haptic	0.690236
silver	visual	0.439678
sizzling	auditory	0.236755
sizzling	visual	0.106354
skinny	visual	0.577205
skinny	visual	0.504878
slick	visual	0.446512
slick	visual	0.445652
slimy	haptic	0.341719
slimy	haptic	0.380631
slippery	haptic	0.467933
slippery	haptic	0.36699
slushy	gustatory	0.262579
slushy	haptic	0.320463
small	visual	0.474304
small	visual	0.550279
smelly	olfactory	0.517073
smelly	olfactory	0.538058
smoky	olfactory	0.222034
smoky	visual	0.316629
smooth	haptic	0.451613
smooth	haptic	0.497423
snarling	auditory	0.660156
snarling	auditory	0.48329
snorting	auditory	0.544833
snorting	auditory	0.63522
soapy	gustatory	0.343669
soapy	visual	0.34749
sodden	haptic	0.36039
sodden	visual	0.339779
soft	haptic	0.436508
soft	auditory	0.821012
solid	haptic	0.403587
solid	haptic	0.402273
sonorous	auditory	0.654711
sonorous	auditory	0.741007
sore	haptic	0.871245
sore	haptic	0.401114
soundless	visual	0.464497
soundless	auditory	0.569721
sour	gustatory	0.556507
sour	gustatory	0.506527
sparkly	visual	0.946188
sparkly	visual	0.31064
speckled	visual	0.468708

filthy	visual	0.371429
filthy	visual	0.321755
flaky	visual	0.268293
flaky	visual	0.351579
flat	visual	0.467626
flat	gustatory	0.39905
fleshy	visual	0.451087
fleshy	gustatory	0.289466
flexible	haptic	0.474453
flexible	haptic	0.503817
flickering	visual	0.654362
flickering	visual	0.808163
floppy	visual	0.567935
floppy	haptic	0.29264
floral	olfactory	0.599315
floral	visual	0.821277
flowery	visual	0.549723
flowery	olfactory	0.755396
fluffy	haptic	0.471526
fluffy	visual	0.284483
foamy	visual	0.32906
foamy	visual	0.360417
foggy	visual	0.654362
foggy	visual	0.487047
forked	visual	0.644689
forked	visual	0.48164
fragrant	olfactory	0.577844
fragrant	olfactory	0.560773
freezing	haptic	0.363458
freezing	haptic	0.420779
fresh	olfactory	0.486486
fresh	olfactory	0.165505
frosty	visual	0.434144
frosty	visual	0.331839
fruity	olfactory	0.471591
fruity	gustatory	0.422658
fuzzy	haptic	0.462075
fuzzy	visual	0.95045
gamy	gustatory	0.247573
gamy	gustatory	0.3361
garlicky	olfactory	0.630573
garlicky	gustatory	0.404711
gigantic	visual	0.366089
gigantic	visual	0.509615

speckled	visual	0.73454
spicy	olfactory	0.573668
spicy	gustatory	0.42887
spiky	haptic	0.479381
spiky	visual	0.453581
spotted	visual	0.92891
spotted	visual	0.536818
square	visual	0.538462
square	visual	0.552279
squeaking	auditory	0.5225
squeaking	auditory	0.70949
squealing	auditory	0.480952
squealing	auditory	0.589385
stagnant	olfactory	0.273871
stagnant	visual	0.475374
stale	gustatory	0.46438
stale	gustatory	0.30767
steep	visual	0.620579
steep	visual	0.529101
stenchy	olfactory	0.554572
stenchy	olfactory	0.477747
sticky	haptic	0.392996
sticky	haptic	0.454756
stinging	haptic	0.484375
stinging	haptic	0.498592
stinky	olfactory	0.383929
stinky	olfactory	0.609091
straight	visual	0.776471
straight	visual	0.77381
striped	visual	0.875648
striped	visual	0.959459
strong	haptic	0.497297
strong	gustatory	0.482838
sturdy	visual	0.50411
sturdy	haptic	0.468665
sunny	visual	0.482234
sunny	visual	0.578947
sweaty	visual	0.29202
sweaty	haptic	0.278195
sweet	olfactory	0.696768
sweet	gustatory	0.465066
swift	visual	0.602305
swift	visual	0.579832
swinging	visual	0.463104

giggling	auditory	0.421171
giggling	auditory	0.466825
glamorous	visual	0.571288
glamorous	visual	0.308285
glistening	visual	0.889831
glistening	visual	0.537433
glittery	visual	0.715356
glittery	visual	0.621019
glossy	visual	0.4725
glossy	visual	0.385776
glowing	visual	0.655455
glowing	visual	0.820084
gold	visual	0.609907
gold	visual	0.682862
gooey	haptic	0.330709
gooey	gustatory	0.243655
gorgeous	visual	0.578171
gorgeous	auditory	0.798354
grainy	visual	0.712727
grainy	haptic	0.330258
granular	visual	0.280353
granular	haptic	0.321596
grassy	gustatory	0.442029
grassy	visual	0.402322
gray	visual	0.955556
gray	visual	0.917749
greasy	gustatory	0.26484
greasy	haptic	0.414188
green	visual	0.643963
green	visual	0.575301
grinding	visual	0.284
grinding	auditory	0.438725
gritty	visual	0.382022
gritty	haptic	0.43377
groaning	auditory	0.571429
groaning	auditory	0.567568
grotesque	visual	0.621795
grotesque	visual	0.508108
growling	auditory	0.479218
growling	auditory	0.632968
gurgling	auditory	0.460396
gurgling	auditory	0.346792
hairy	haptic	0.414416
hairy	visual	0.5025
swinging	auditory	0.656667
tall	visual	0.620991
tall	visual	0.808765
tangerine	visual	0.769912
tangerine	visual	0.361478
tangy	gustatory	0.477064
tangy	gustatory	0.475128
tapering	visual	0.331461
tapering	visual	0.544118
tarry	visual	0.32636
tarry	visual	0.416021
tart	gustatory	0.474201
tart	gustatory	0.508929
tasteless	visual	0.421725
tasteless	gustatory	0.553314
tender	haptic	0.553009
tender	gustatory	0.353896
tepid	haptic	0.461538
tepid	gustatory	0.365285
thorny	haptic	0.473538
thorny	haptic	0.488312
thudding	auditory	0.539474
thudding	auditory	0.411622
thumping	auditory	0.449883
thumping	haptic	0.538482
ticklish	haptic	0.607455
ticklish	haptic	0.58104
tight	haptic	0.526455
tight	haptic	0.486874
tinkling	auditory	0.356796
tinkling	auditory	0.408163
tiny	visual	0.585434
tiny	auditory	0.819328
tough	haptic	0.483791
tough	gustatory	0.361419
translucent	visual	0.706678
translucent	visual	0.854251
transparent	visual	0.814672
transparent	visual	0.834008
triangular	visual	0.621455
triangular	visual	0.42723
ugly	visual	0.598071
ugly	visual	0.533791
uneven	visual	0.654545

handsome	visual	0.679739
handsome	visual	0.613095
happy	visual	0.589286
happy	auditory	0.712687
hard	gustatory	0.407328
hard	haptic	0.433708
harsh	auditory	0.521909
harsh	haptic	0.336207
heavy	visual	0.474801
heavy	visual	0.428635
herby	gustatory	0.422222
herby	visual	0.330693
high	auditory	0.791165
high	visual	0.758364
hissing	auditory	0.5
hissing	auditory	0.483627
hoarse	auditory	0.709821
hoarse	auditory	0.459732
hollow	visual	0.4197
hollow	visual	0.54388
honeyed	gustatory	0.342056
honeyed	auditory	0.773333
hot	gustatory	0.38758
hot	haptic	0.366667
howling	auditory	0.488943
howling	auditory	0.548649
huge	visual	0.382269
huge	visual	0.562147
humid	haptic	0.349765
humid	haptic	0.307571
humming	auditory	0.603499
humming	auditory	0.755304
hushed	auditory	0.529915
hushed	auditory	0.615385
husky	haptic	0.385765
husky	auditory	0.889831
icy	haptic	0.563025

uneven	haptic	0.456311
unripe	gustatory	0.277448
unripe	gustatory	0.34188
vinegary	olfactory	0.353333
vinegary	gustatory	0.486413
vivid	visual	0.901408
vivid	visual	0.281095
wailing	auditory	0.494186
wailing	auditory	0.633333
warbling	auditory	0.706806
warbling	auditory	0.572368
warm	haptic	0.429671
warm	haptic	0.557065
waxy	haptic	0.432373
waxy	haptic	0.439589
weak	haptic	0.562937
weak	visual	0.864035
weightless	visual	0.672535
weightless	haptic	0.491573
wet	visual	0.343685
wet	haptic	0.327451
whining	auditory	0.568915
whining	auditory	0.502347
whistling	auditory	0.561497
whistling	auditory	0.578804
white	visual	0.788
white	visual	0.889868
wide	visual	0.540616
wide	visual	0.666667
wiry	visual	0.546125
wiry	haptic	0.436474
wispy	visual	0.792579
wispy	visual	0.416058
woolly	haptic	0.517857
woolly	haptic	0.486811
yellow	visual	0.816794
yellow	visual	0.849206

List of 387 Adjectives and their Sensory Values for each corresponding Representational System, which across both of the modalities equals 774 words.

Table 35: Summary of the findings of Argamon *et al.'s* (2003) gender study

Pronouns	Tag	Female μ ± stderr	Male μ ± stderr	t-test	Female median	Male median
he	M	271 ± 9 3	305 ± 11	p<0.05 *	276	305
her	F	53.8 ± 5.1	18.5 ± 3.5	p<0.0001	29.8	5.60
hers	F	53.8 ± 5.1	18.5 ± 3.5	p<0.0001	29.8	5.60
herself	F	53.8 ± 5.1	18.5 ± 3.5	p<0.0001	29.8	5.60
him	M	271 ± 9 3	305 ± 11	p<0.05 *	276	305
himself	M	271 ± 9 3	305 ± 11	p<0.05 *	276	305
his	M	271 ± 9 3	305 ± 11	p<0.05 *	276	305
I	F	149 ± 14	86 ± 8	p<0.0002	66.7	50.2
it	F	89.1 ± 2.8	86.7 ± 2.4	**n/s**	85.3	82.9
its	M	15.3 ± 0.93	19.0 ± 0.79	p<0.005	12.2	19.0
me	F	149 ± 14	86 ± 8	p<0.0002	66.7	50.2
mine	F	149 ± 14	86 ± 8	p<0.0002	66.7	50.2
my	F	149 ± 14	86 ± 8	p<0.0002	66.7	50.2
myself	F	149 ± 14	86 ± 8	p<0.0002	66.7	50.2
our	F	149 ± 14	86 ± 8	p<0.0002	66.7	50.2
ours	F	149 ± 14	86 ± 8	p<0.0002	66.7	50.2
ourselves	F	149 ± 14	86 ± 8	p<0.0002	66.7	50.2
she	F	53.8 ± 5.1	18.5 ± 3.5	p<0.0001	29.8	5.60
their	F	97.8 ± 4.6	81.8 ± 2.7	p<0.005	83.9	78.8
theirs	F	97.8 ± 4.6	81.8 ± 2.7	p<0.005	83.9	78.8
them	F	97.8 ± 4.6	81.8 ± 2.7	p<0.005	83.9	78.8
themselves	F	97.8 ± 4.6	81.8 ± 2.7	p<0.005	83.9	78.8
they	F	97.8 ± 4.6	81.8 ± 2.7	p<0.005	83.9	78.8
us	F	149 ± 14	86 ± 8	p<0.0002	66.7	50.2
we	F	149 ± 14	86 ± 8	p<0.0002	66.7	50.2
You	F	63.9 ± 8.0	30.0 ± 5.2	p<0.0005	16.7	3.9
Your	F	63.9 ± 8.0	30.0 ± 5.2	p<0.0005	16.7	3.9
Yours	F	63.9 ± 8.0	30.0 ± 5.2	p<0.0005	16.7	3.9
Yourself	F	63.9 ± 8.0	30.0 ± 5.2	p<0.0005	16.7	3.9

Table 36: Shakespeare, Marlowe and Carey's Works and how they were broken into chunks

ID	YEAR*	TITLE	TYPE	SHORT TITLE	IN WORK
WILLIAM SHAKESPEARE					
1	1589	Comedy of Errors	C	C1	Comedy of Errors
2	1590	Henry VI, Part II	H	H1	Henry VI, Part II
3	1590	Henry VI, Part III	H	H2	Henry VI, Part III
4	1591	Henry VI, Part I	H	H3	Henry VI, Part I
5	1592	Richard III	H	H4	Richard III
6	1593	Taming of the Shrew	C	C2	Taming of the Shrew
7	1593	Titus Andronicus	T	T1	Titus Andronicus
8	1593	Venus and Adonis	P	P1	Venus and Adonis
9	1594	Love's Labour's Lost	C	C4	Love's Labour's Lost
10	1594	Romeo and Juliet	T	T2	Romeo and Juliet
11	1594	The Rape of Lucrece	P	P2	The Rape of Lucrece
12	1594	Two Gentlemen of Verona	C	C3	Two Gentlemen of Verona

13	1595	Midsummer Night's Dream	C	C5	Midsummer Night's Dream
14	1595	Richard II	H	H5	Richard II
15	1596	King John	H	H6	King John
16	1596	Merchant of Venice	C	C6	Merchant of Venice
17	1597	Henry IV, Part I	H	H7	Henry IV, Part I
18	1597	Henry IV, Part II	H	H8	Henry IV, Part II
19	1598	Henry V	H	H9	Henry V
20	1598	Much Ado about Nothing	C	C7	Much Ado about Nothing
21	1599	As You Like It	C	C9	As You Like It
22	1599	Julius Caesar	T	T3	Julius Caesar
23	1599	Love's Answer	P	P5	The Passionate Pilgrim
24	1599	Sonnets to sundry notes of music	P	P4	The Passionate Pilgrim
25	1599	The Passionate Pilgrim	P	P3	The Passionate Pilgrim
26	1599	Twelfth Night	C	C8	Twelfth Night
27	1600	Hamlet	T	T4	Hamlet
28	1600	Merry Wives of Windsor	C	C10	Merry Wives of Windsor
29	1601	The Phoenix and the Turtle	P	P6	The Phoenix and the Turtle
30	1601	Threnos	P	P7	The Phoenix and the Turtle
31	1601	Troilus and Cressida	C	C11	Troilus and Cressida
32	1602	All's Well That Ends Well	C	C12	All's Well That Ends Well
33	1604	Measure for Measure	C	C13	Measure for Measure
34	1604	Othello	T	T5	Othello
35	1605	King Lear	T	T6	King Lear
36	1605	Macbeth	T	T7	Macbeth
37	1606	Anthony and Cleopatra	T	T10	Anthony and Cleopatra
38	1607	Coriolanus	T	T8	Coriolanus
39	1607	Timon of Athens	T	T9	Timon of Athens
40	1608	Pericles	C	C14	Pericles
41	1609	A Lover's Complaint	P	P8	The Passionate Pilgrim
42	1609	Cymbeline	C	C15	Cymbeline
43	1609	Sonnets	P	P9	Sonnets
44	1610	Winter's Tale	C	C16	Winter's Tale
45	1611	Tempest	C	C17	Tempest
46	1612	Henry VIII	H	H10	Henry VIII
CHRISTOPHER MARLOWE					
47	1590	Tamburlaine Part I		M1	Tamburlaine The Great Part I
48	1590	Tamburlaine Part II		M2	Tamburlaine The Great Part II
49		Edward II	H	M3	Edward II
50		The Jew of Malta	T	M4	The Jew of Malta
51		Doctor Faustus		M5	Doctor Faustus

52		Dido Queen of Carthage		M6	Dido Queen of Carthage
53		The Massacre at Paris		M7	The Massacre at Paris with the Death of the Duke of Guise
54		Hero and Leander	P	M8	Hero and Leander
55		The Passionate Shepherd	P	M9	The Passionate Shepherd to His Love
56		Walter Raleigh	P	M10	The Passionate Shepherd to His Love
ELIZABETH CAREY					
57	1612	The Tragedy of Mariam	T	EC1	The Tragedy of Mariam, the Fair Queen of Jewry

Type: C = Comedies, H = Histories, T = Tragedies, P = Poems

* The Year may not have any bearing as many works may well have been written earlier. In Marlowe's case, all but two of his works were published after his death.

Table 37: The list of the poems by Shakespeare, Barnfield, Griffin, Marlowe including the 12 unknown authored poems in The Passionate Pilgrim Poems by Author and Abbreviated ID.

ID	Abbreviated	Author
1	1S	William Shakespeare
2	2S	William Shakespeare
3	3S	William Shakespeare
4	4U	Unknown
5	5S	William Shakespeare
6	6U	Unknown
7	7U	Unknown
8	8B	Richard Barnfield
9	9U	Unknown
10	10U	Unknown
11	11G	Bartholomew Griffin
12	12U	Unknown (Thomas Deloney)
13	13U	Unknown
14	14U	Unknown
15	15U	Unknown
16	16U	Unknown
17	17S	William Shakespeare
18	18U	Unknown
19	19U	Unknown
20	20M	Christopher Marlowe and Walter Raleigh
21	21B	Richard Barnfield

Table 38: Pearson correlation coefficient, R, results of RPAS, the five sensory elements (VAHOG), and the four Referential Activity Power elements.

Correlations

		R	P	A	S	
Richness (R)	Pearson Correlation	1	.399**	-.833**	.456**	
	Sig. (2-tailed)		0.002	0	0	
	N	57	57	57	57	
Personal_Pronouns (P)	Pearson Correlation	.399**	1	-.451**	.366**	
	Sig. (2-tailed)	0.002		0	0.005	
	N	57	57	57	57	
RA Power (A)	Pearson Correlation	-.833**	-.451**	1	-.575**	
	Sig. (2-tailed)	0	0		0	
	N	57	57	57	57	
Sensory (S)	Pearson Correlation	.456**	.366**	-.575**	1	
	Sig. (2-tailed)	0	0.005	0		
	N	57	57	57	57	
		V	A	H	O	G
Sensory - Visual (V)	Pearson Correlation	1	.284*	.715**	.784**	.571**
	Sig. (2-tailed)		0.032	0	0	0
	N	57	57	57	57	57
Sensory - Auditory (A)	Pearson Correlation	.284*	1	-0.038	0.167	-0.119
	Sig. (2-tailed)	0.032		0.777	0.215	0.378
	N	57	57	57	57	57
Sensory - Haptic (H)	Pearson Correlation	.715**	-0.038	1	.632**	.772**
	Sig. (2-tailed)	0	0.777		0	0
	N	57	57	57	57	57
Sensory - Olfactory (O)	Pearson Correlation	.784**	0.167	.632**	1	.628**
	Sig. (2-tailed)	0	0.215	0		0
	N	57	57	57	57	57
Sensory - Gustatory (G)	Pearson Correlation	.571**	-0.119	-0.119	.628**	1
	Sig. (2-tailed)	0	0.378	0.378	0	
	N	57	57	57	57	57
		A	C	P	PRON	
RA Power - Article (A)	Pearson Correlation	1	.800**	.899**	.686**	
	Sig. (2-tailed)		0	0	0	
	N	57	57	57	57	
RA Power - Conjunctive (C)	Pearson Correlation	.800**	1	.859**	.563**	
	Sig. (2-tailed)	0		0	0	
	N	57	57	57	57	
RA Power - Preposition (P)	Pearson Correlation	.899**	.859**	1	.706**	
	Sig. (2-tailed)	0	0		0	
	N	57	57	57	57	

	Pearson Correlation	.686**	.563**	.706**	1	
RA Power - Pronoun (PRON)	Sig. (2-tailed)	0	0	0		
	N	57	57	57	57	

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Table 39: PCA Descriptive Statistics Shakespeare, Marlowe, and Cary

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
Richness	2.2034958792E1	1.52579825997E1	57
Gender Pronouns	5.9897442035E0	2.02163110795E1	57
RA Score	4.2469364069E1	2.09366559396E1	57
Auditory	.1469911971	1.13237041280E-1	57
Gustatory	.1338646183	9.37442724914E-2	57
Haptic	.1317616465	8.45222795429E-2	57
Olfactory	.1353649594	1.20863535652E-1	57
Visual	.7033064182	3.70917467781E-1	57

Table 40: PCA Correlation Matrix for Shakespeare, Marlowe and Cary

Correlation Matrix[a]

		Richness	P_Pronouns	RA Score	Auditory	Gustatory	Haptic	Olfactory	Visual
Correlation	Richness	1.000	.399	-.833	.606	.311	.195	.296	.372
	P_Pronouns	.399	1.000	-.451	.252	.210	.340	.169	.362
	RA Score	-.833	-.451	1.000	-.439	-.430	-.343	-.430	-.520
	Auditory	.606	.252	-.439	1.000	-.119	-.038	.167	.284
	Gustatory	.311	.210	-.430	-.119	1.000	.772	.628	.571
	Haptic	.195	.340	-.343	-.038	.772	1.000	.632	.715
	Olfactory	.296	.169	-.430	.167	.628	.632	1.000	.784
	Visual	.372	.362	-.520	.284	.571	.715	.784	1.000
Sig. (1-tailed)	Richness		.001	.000	.000	.009	.073	.013	.002
	P_Pronouns	.001		.000	.029	.058	.005	.104	.003
	RA Score	.000	.000		.000	.000	.005	.000	.000
	Auditory	.000	.029	.000		.189	.389	.107	.016
	Gustatory	.009	.058	.000	.189		.000	.000	.000
	Haptic	.073	.005	.005	.389	.000		.000	.000
	Olfactory	.013	.104	.000	.107	.000	.000		.000
	Visual	.002	.003	.000	.016	.000	.000	.000	

a. Determinant = .004

Table 41: PCA KMO and Bartlett's Test for Shakespeare, Marlowe, and Cary

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.722
Bartlett's Test of Sphericity	Approx. Chi-Square	290.851
	df	28
	Sig.	.000

Table 42: PCA Communalities for Shakespeare, Marlowe, and Cary

Communalities

	Initial	Extraction
Richness	1.000	.832
Gender Pronouns	1.000	.354
RA Score	1.000	.787
Auditory	1.000	.695
Gustatory	1.000	.772
Haptic	1.000	.830
Olfactory	1.000	.711
Visual	1.000	.770

Extraction Method: Principal Component Analysis.

Table 43: PCA Total Variance for Shakespeare, Marlowe and Cary

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.965	49.565	49.565	3.965	49.565	49.565	3.163	39.537	39.537
2	1.786	22.322	71.887	1.786	22.322	71.887	2.588	32.350	71.887
3	.807	10.093	81.980						
4	.652	8.148	90.128						
5	.327	4.084	94.213						
6	.221	2.766	96.979						
7	.126	1.581	98.560						
8	.115	1.440	100.000						

Extraction Method: Principal Component Analysis.

Table 44: PCA Component Matrix for Shakespeare, Marlowe, and Cary

Component Matrix[a]

	Component	
	1	2
Richness	.684	-.603
Gender Pronouns	.532	
RA Score	-.787	.409
Auditory	.388	-.737
Gustatory	.734	.483
Haptic	.755	.510
Olfactory	.777	.327
Visual	.855	

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Table 45: PCA Rotated Component matrix for Shakespeare, Marlowe, and Cary

Rotated Component Matrix[a]

	Component	
	1	2
Richness		.895
Gender Pronouns		.535
RA Score	-.377	-.803
Auditory		.822
Gustatory	.877	
Haptic	.910	
Olfactory	.817	
Visual	.799	.363

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Figure 36: PCA Scree Plot for Shakespeare, Marlowe, and Cary

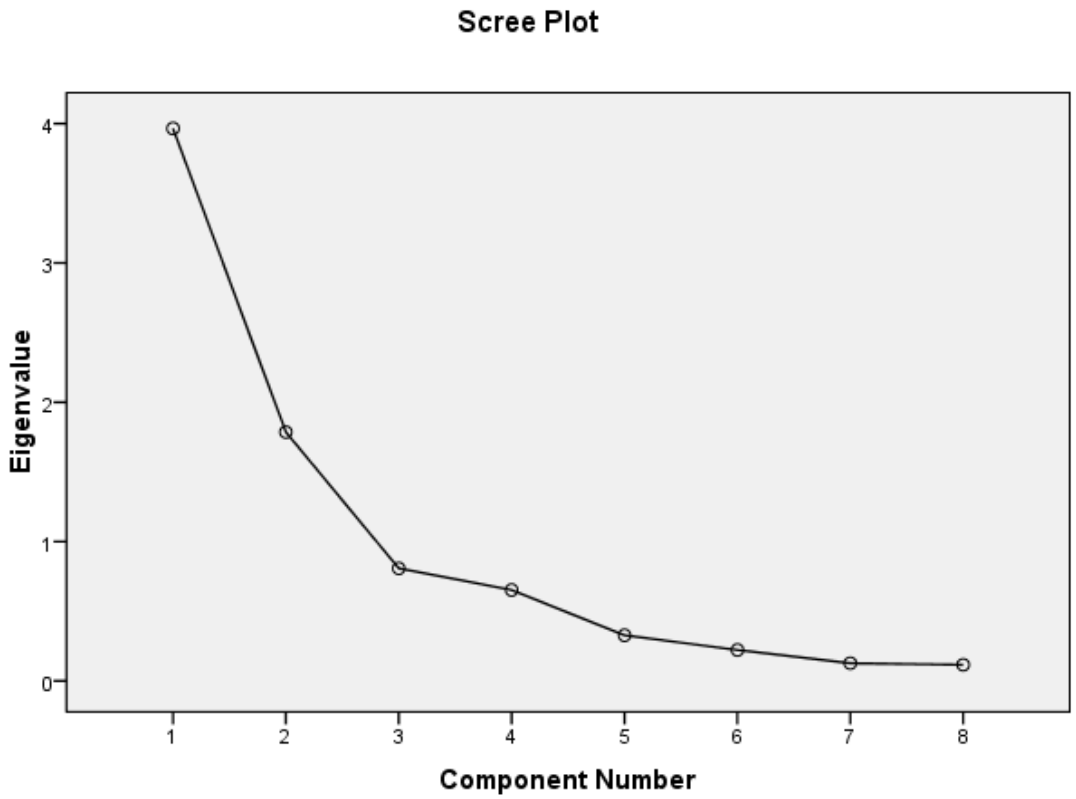


Table 46: LDA Eigenvalues of the first two canonical functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	2.266[a]	79.1	79.1	.833
2	.598[a]	20.9	100.0	.612

a. First 2 canonical discriminant functions were used in the analysis.

Table 47: LDA Wilks' Lambda results of the two canonical functions

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.192	66.919	16	.000
2	.626	18.985	7	.008

Table 48: LDA Discriminant Function coefficients for the two canonical functions

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
Richness	1.115	.092
Gender	.369	-.346
RA Score	-.119	.832
Auditory	.816	.394
Gustatory	-.084	.672
Haptic	-1.733	-.127
Olfactory	.642	-.251
Visual	-.180	.831

Table 49: LDA Group Centroids of the three Playwrights for both canonical functions

Functions at Group Centroids

Playwrights	Function	
	1	2
1	-.645	.070
2	3.423	.327
3	1.205	-5.037

Unstandardized canonical discriminant functions evaluated at group means

Figure 37: The first two dimensions of a canonical discriminant analysis applied to the uncontested works of Shakespeare, Marlowe, and Cary

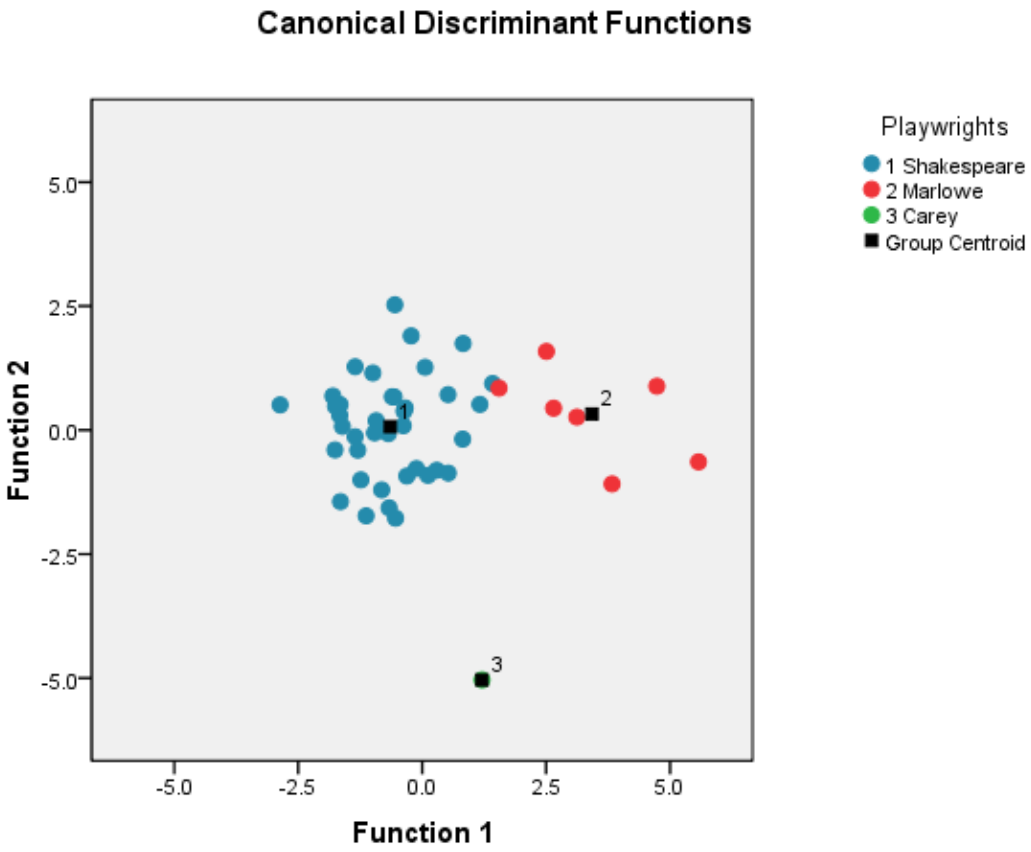


Figure 38: PtoR discriminates chunks 6, 8, 10, 15, 16, 17, and 18 mainly by Richness, while the RAtoR inset Radar plot highlights a constant Referential Activity Power plot in the center (green). Of note, the four Shakespearian clusters marked with a red circle are those commonly attributed to Shakespeare. Further, the ellipses are our visual clustering assignment.

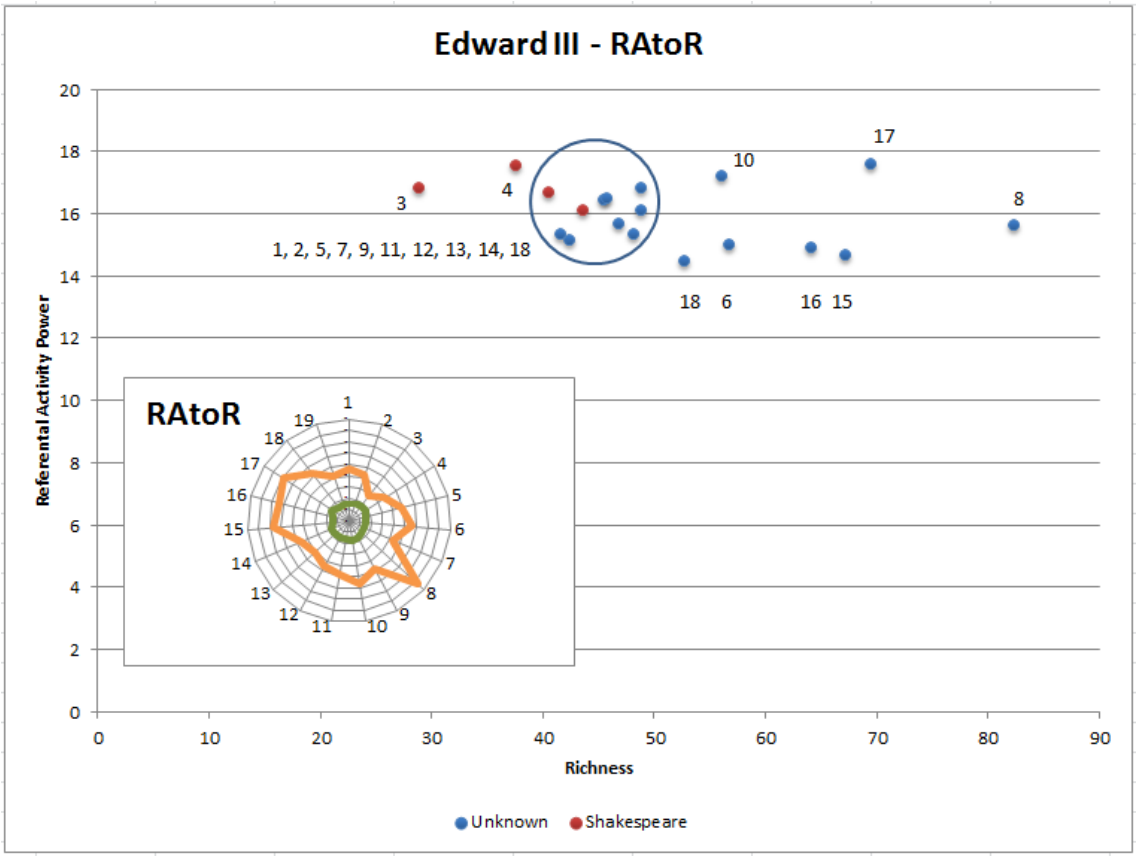


Table 50: Iris Murdoch's novels by year published

ID	Iris Murdoch Novels	Published
B1	Under the Net	1954
B2	The Flight from the Enchanter	1956
B3	The Sandcastle	1957
B4	The Bell	1958
B5	A Severed Head	1961
B6	An Unofficial Rose	1962
B7	The Unicorn	1963
B8	The Italian Girl	1964
B9	The Red and the Green	1965
B10	The Time of the Angels	1966
B11	The Nice and the Good	1968
B12	Bruno's Dream	1969
B13	A Fairly Honourable Defeat	1970
B14	An Accidental Man	1971
B15	The Black Prince	1973
B16	The Sacred and Profane Love Machine	1974
B17	A Word Child	1975
B18	Henry and Cato	1976
B19	The Sea, the Sea	1978
B20	Nuns and Soldiers	1980
B21	The Philosopher's Pupil	1983
B22	The Good Apprentice	1985
B23	The Book and the Brotherhood	1987
B24	The Message to the Planet	1989
B25	The Green Knight	1993
B26	Jackson's Dilemma	1995

Table 51: P.D. James' novels by year published

ID	P.D. James Novels	Published
B1	Cover Her Face	1962
B2	A Mind to Murder	1963
B3	Unnatural Causes	1967
B4	Shroud for a Nightingale	1971
B5	An Unsuitable Job for a Woman	1972
B6	The Black Tower	1975
B7	Death of an Expert Witness	1977
B8	Innocent Blood	1980
B9	The Skull Beneath the Skin	1982
B10	A Taste for Death	1986
B11	Devices and Desires	1989
B12	The Children of Men	1992
B13	Original Sin	1994
B14	A Certain Justice	1997
B15	Death in Holy Orders	2001
B16	The Murder Room	2003
B17	The Lighthouse	2005
B18	The Private Patient	2008
B19	Death Comes to Pemberley	2011

Table 52: Iris Murdoch Sensory Word Mann-Whitney U-test 12 year Ranks

Ranks

	Group	N	Mean Rank	Sum of Ranks
Rank of Sensory by Group	1	20	15.60	312.00
	2	6	6.50	39.00
	Total	26		

Table 53: Iris Murdoch Sensory Word Mann-Whitney U-test 12-year statistics

Test Statistics[b]

	Rank of Sensory by Group
Mann-Whitney U	18.000
Wilcoxon W	39.000
Z	-2.559
Asymp. Sig. (2-tailed)	.011
Exact Sig. [2*(1-tailed Sig.)]	.009[a]

a. Not corrected for ties.

b. Grouping Variable: Group

Table 54: P.D. James Sensory Word Mann-Whitney U-test 12 year Ranks

Ranks

	Group	N	Mean Rank	Sum of Ranks
Rank of Sensory	1	15	11.53	173.00
	2	4	4.25	17.00
	Total	19		

Table 55: P.D. James Sensory Word Mann-Whitney U-test 12-year statistics

Test Statistics[b]

	Rank of Sensory
Mann-Whitney U	7.000
Wilcoxon W	17.000
Z	-2.300
Asymp. Sig. (2-tailed)	.021
Exact Sig. [2*(1-tailed Sig.)]	.020[a]

a. Not corrected for ties.

b. Grouping Variable: Group

Figure 39: Iris Murdoch Content Words POS Mean with Standard Error bars. We see the aggregated Content Words part-of-speech is higher than the earlier work for the 12 year period before the diagnosis of AD, and there is no overlap between the Standard Error means. There is more variability in the 12 Years period (25.6 versus 18.2).

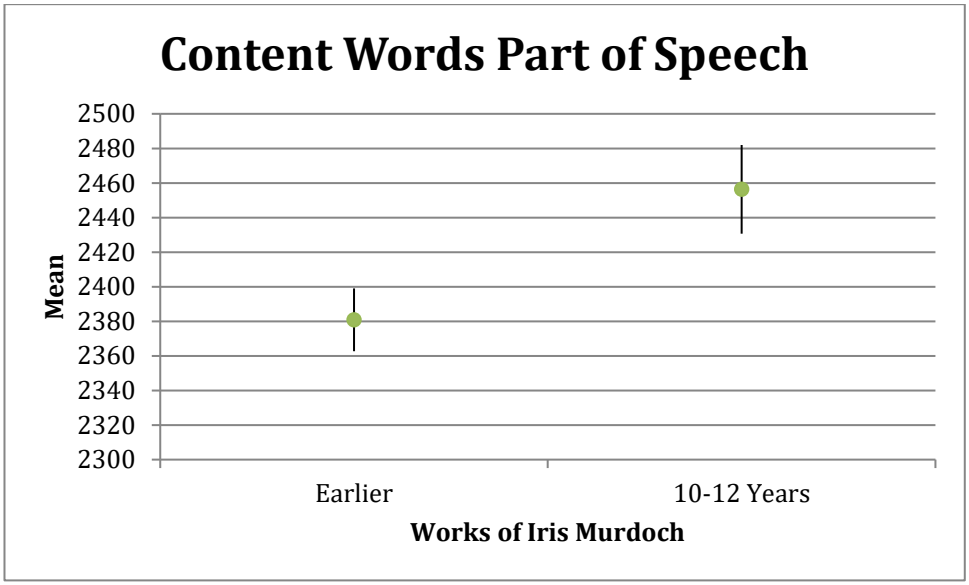


Figure 40: Iris Murdoch Function Words POS Mean with Standard Error bars. We see the aggregated Function Words part-of-speech is part-of-speech is lower than the earlier work for the 12 year period before the diagnosis of AD, and there is no overlap between the Standard Error means. There is more variability in the 12 Years period (25.6 versus 18.2).

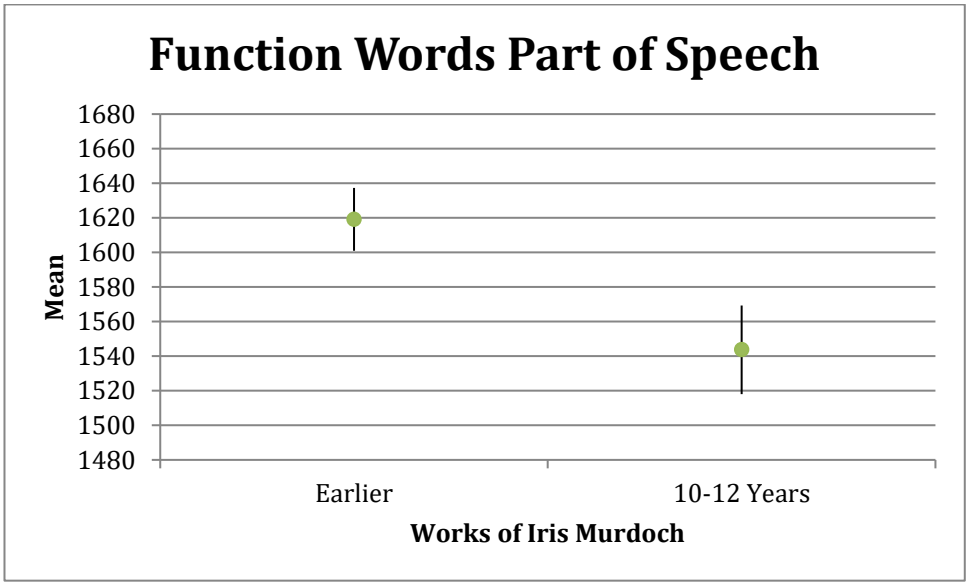


Figure 41: P.D. James Content Words POS Mean with Standard Error bars. We see the aggregated Content Words part-of-speech is higher than the earlier work for the 10-12 year period before death, and there is no overlap between the Standard Error means. There is less variability in the 10-12 Years period (13.28 versus 14.81).

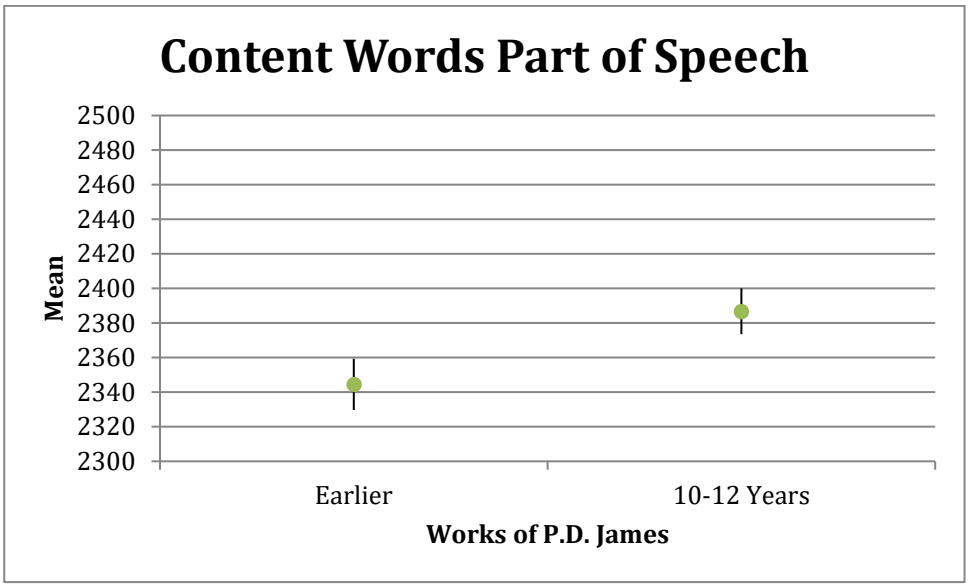


Figure 42: P.D. James Function Words POS Mean with Standard Error bars. We see the aggregated Function Words part-of-speech is lower than the earlier work for the 10-12 year period before the diagnosis of AD, and there is no overlap between the Standard Error means. There is less variability in the 10-12 Years period (13.28 versus 14.81).

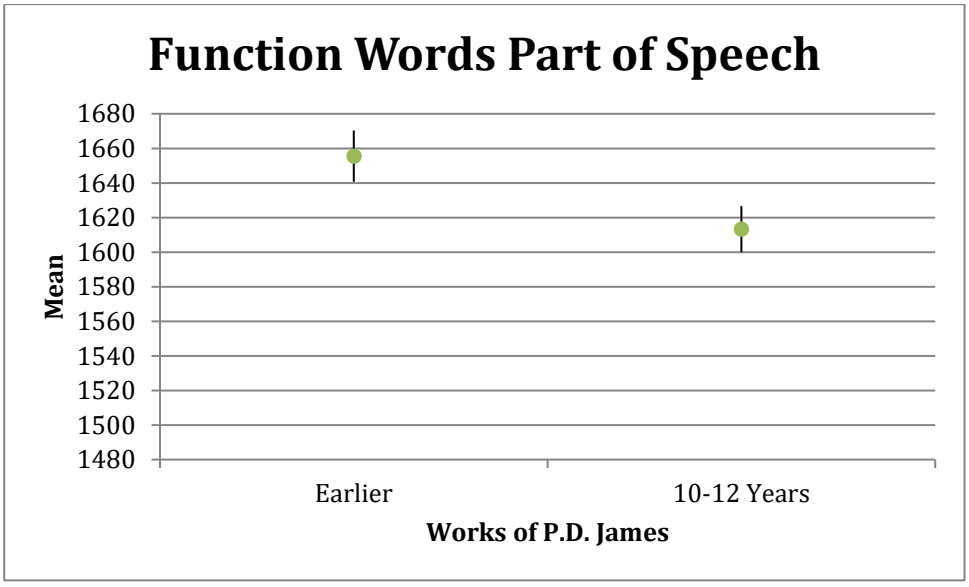


Figure 43: Iris Murdoch Visual Sensory Mean with Standard Error bars

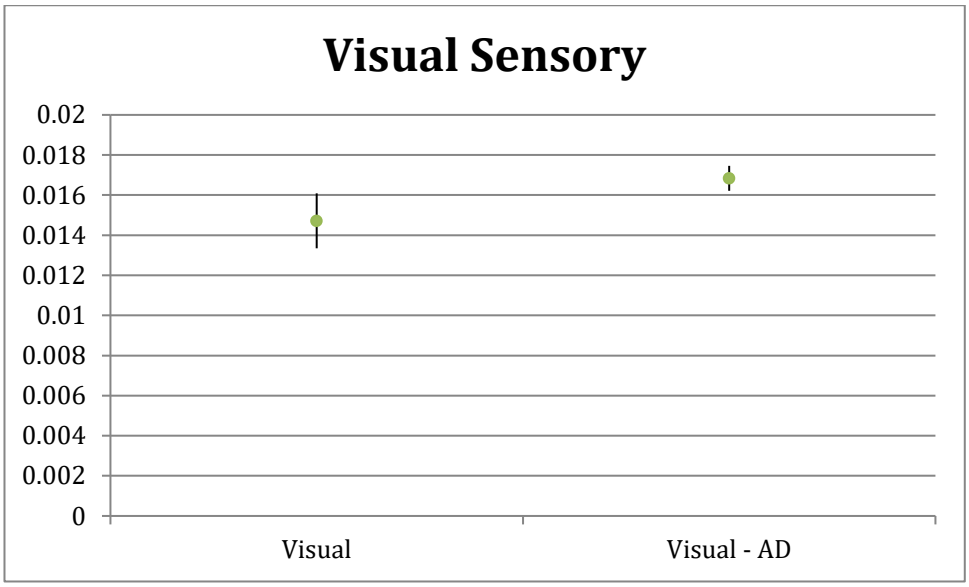


Figure 44: Iris Murdoch Auditory Sensory Mean with Standard error bars

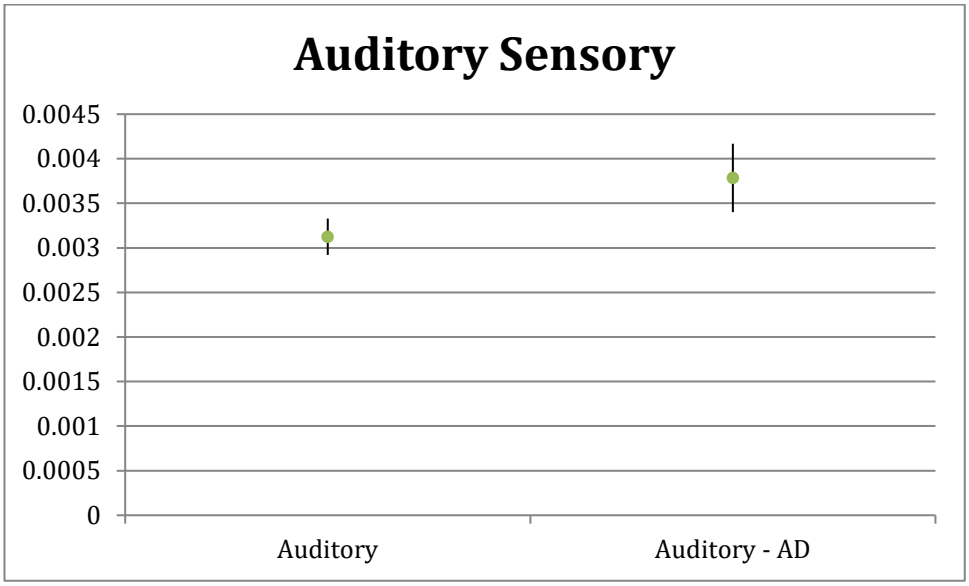


Figure 45: Iris Murdoch Haptic Sensory Mean with Standard Error bars

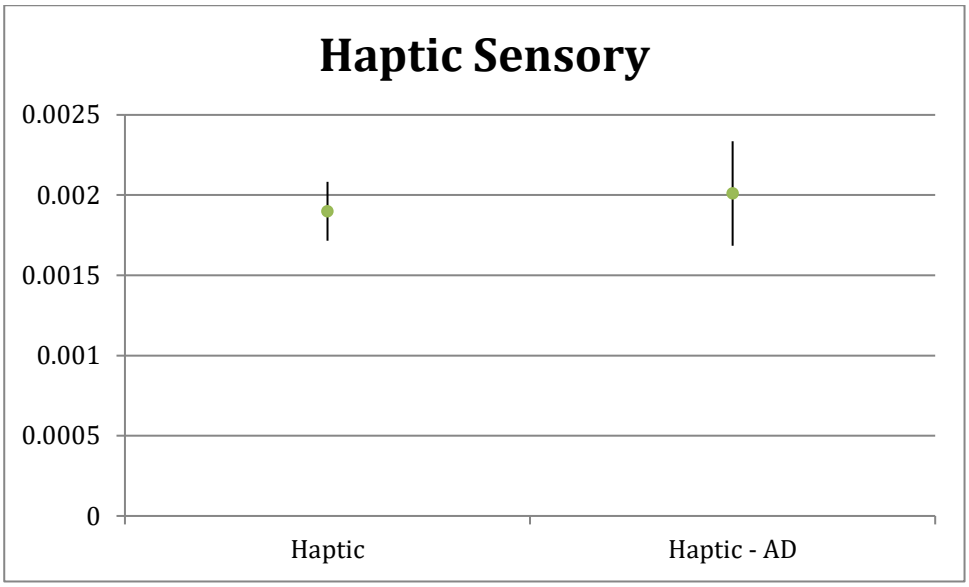


Figure 46: Iris Murdoch Olfactory Sensory Mean with Standard Error bars

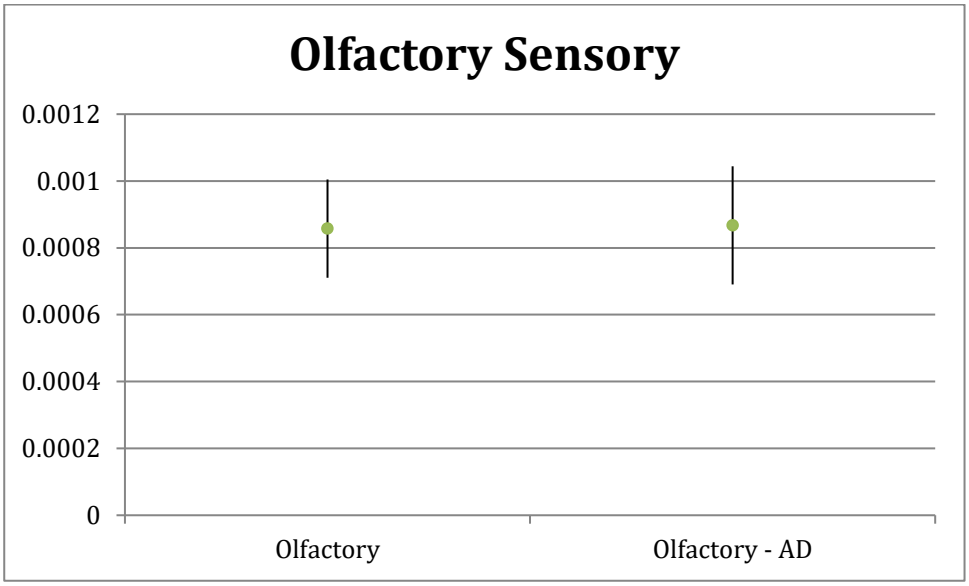


Figure 47: Iris Murdoch Gustatory Sensory Mean with Standard Error bars

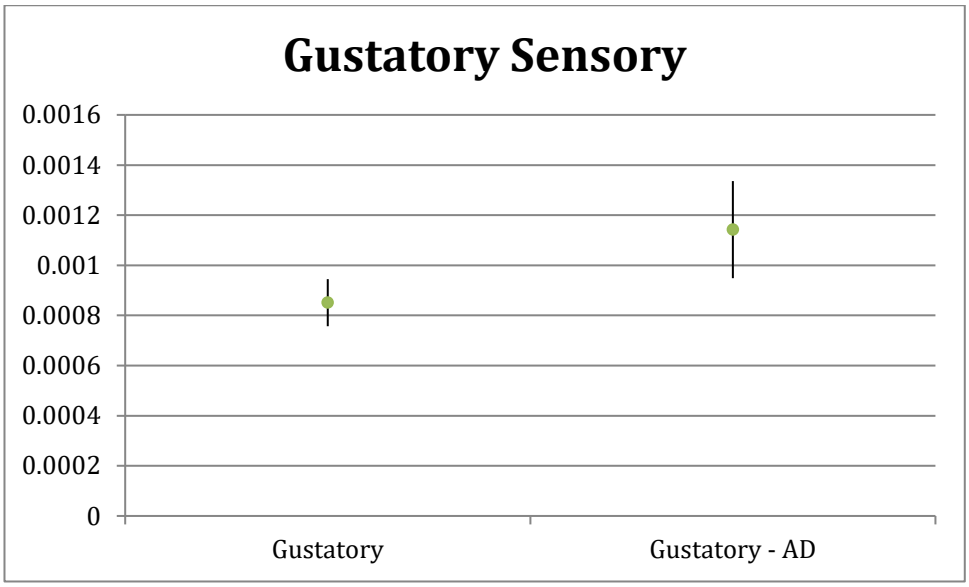


Figure 48: P.D. James Visual Sensory Mean with Standard Error bars

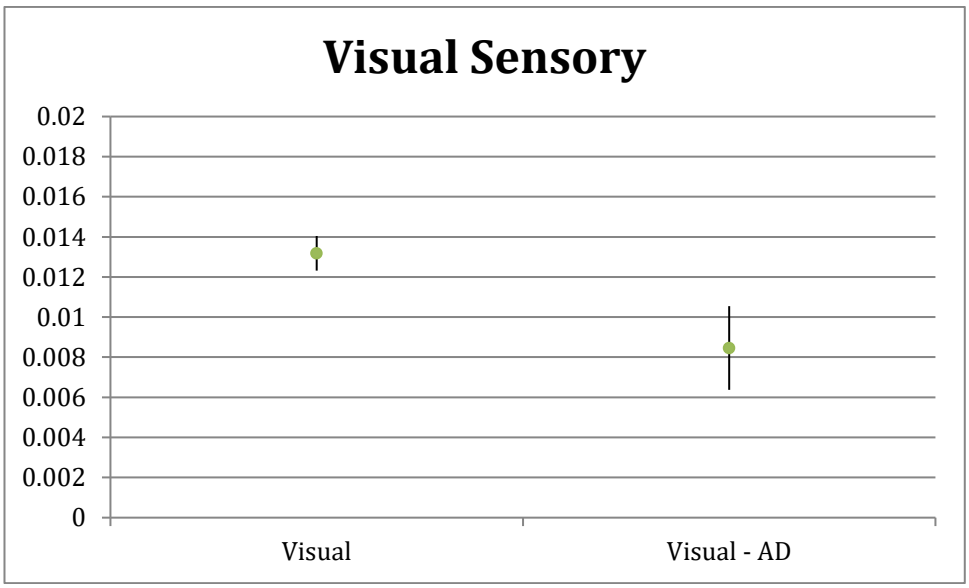


Figure 49: P.D. James Auditory Sensory Mean with Standard Error bars

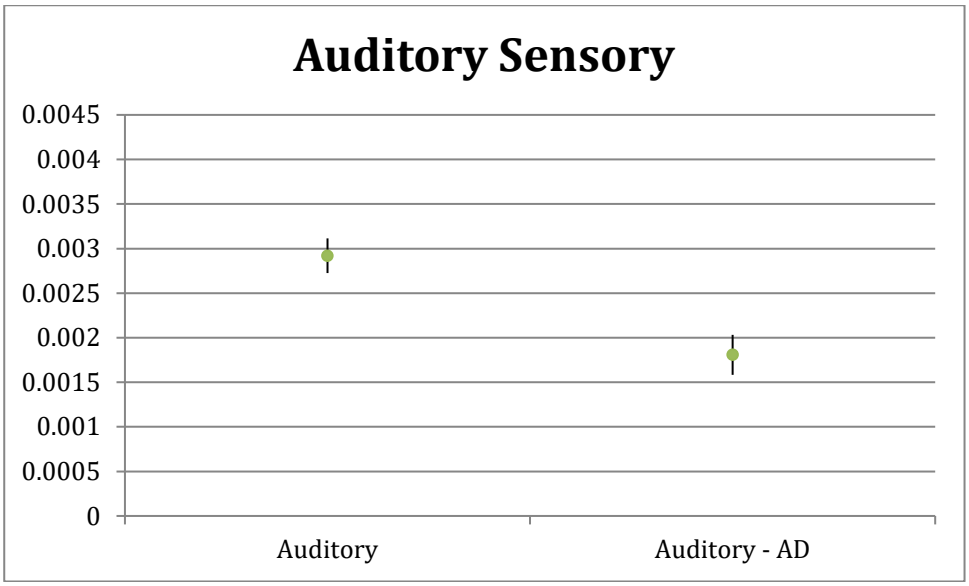


Figure 50: P.D. James Haptic Sensory Mean with Standard Error bars

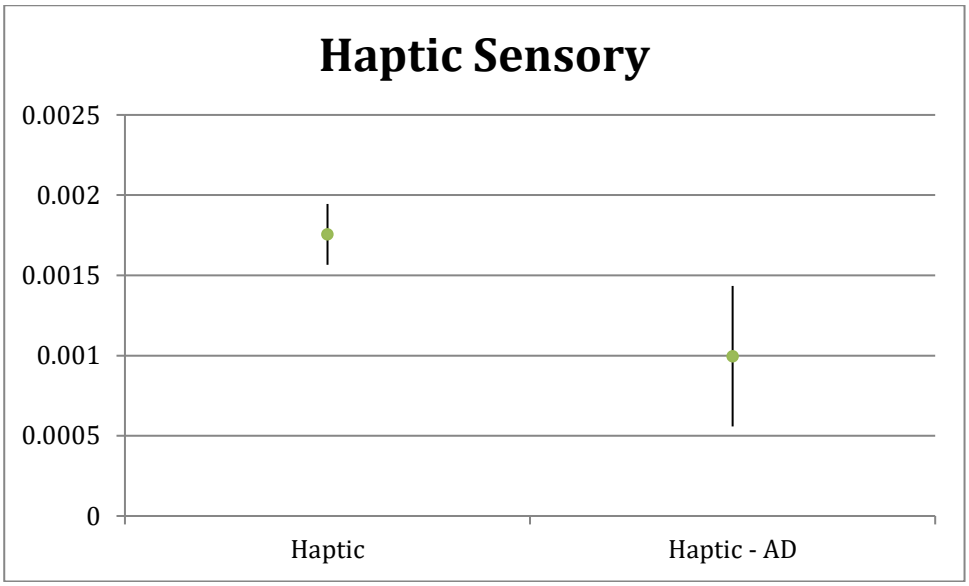


Figure 51: P.D. James Olfactory Sensory Mean with Standard Error bars

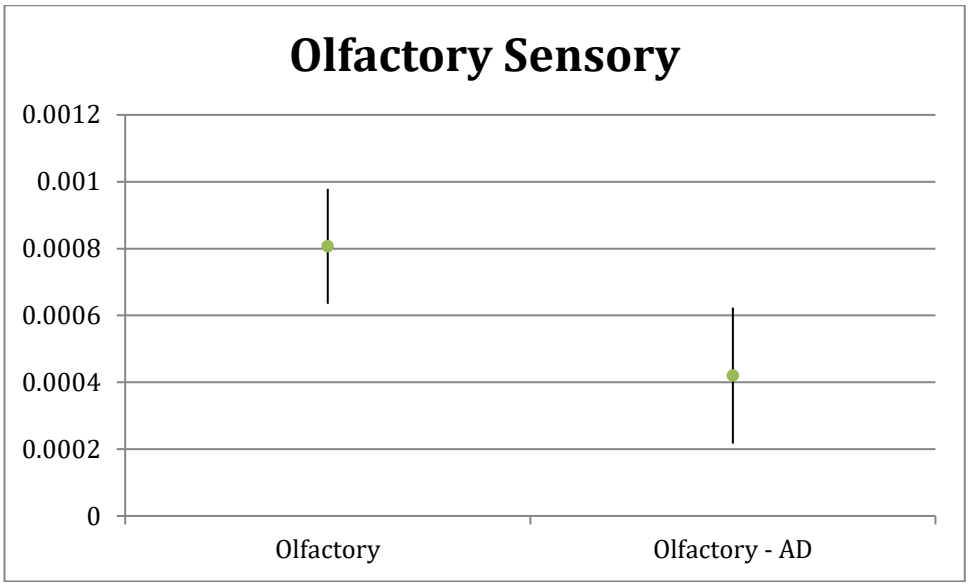


Figure 52: P.D. James Gustatory Sensory Mean with Standard Error bars

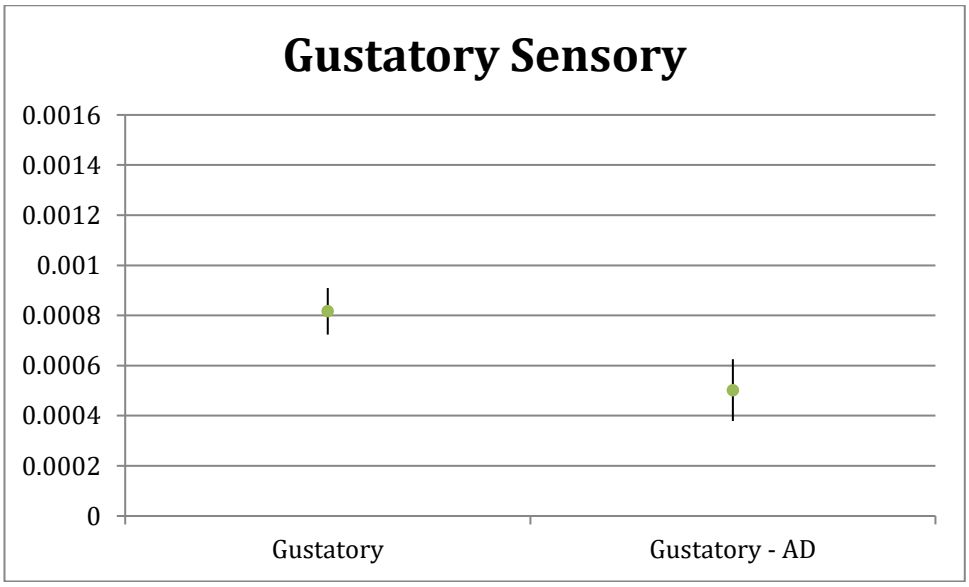


Table 56: RPAV values of Iris Murdoch's novels

Work	Richness	Gender	RA Power	Auditory	Gustatory	Haptic	Olfactory	Visual	Sensory
B1	26.775	.2274041057	14.64112357	0.001893795	0.000606055	0.002230414	0.000755071	0.010487843	0.015973177
B2	28.225	.9988632278	10.91681071	0.004351909	0.000934004	0.001351208	0.000697118	0.01319293	0.020527169
B3	27.05	.6818886726	9.404984821	0.002329807	0.000758867	0.001059749	0	0.007099236	0.011247658
B4	29.975	.9992511049	10.688015	0.002438754	0.001595562	0.001201171	0.000433125	0.009126269	0.014794881
B5	27.05	.7374709587	9.552148929	0.002960395	0.000762951	0.002010495	0.001478563	0.016298165	0.02351057
B6	27.95	.9977307804	8.272718929	0.003540101	0.000393781	0.001799469	0.00043903	0.016143835	0.022316217
B7	29.875	.9783518633	11.85225607	0.00423449	0.000879715	0.002699774	0.001005129	0.021633528	0.030452636
B8	26.65	.00607305159	10.66224143	0.003232946	0.000580226	0.002330373	0.002396304	0.018234171	0.02677402
B9	32.85	.6237562588	13.711175	0.002032694	0.000502611	0.001055165	0.000617499	0.01099064	0.015198609
B10	30.5	.9549280115	10.11495821	0.002555026	0.001420173	0.002804278	0.001504859	0.017856794	0.02614113
B11	30.125	.3714940843	13.48681929	0.003345157	0.000711792	0.002249554	0.000258933	0.01959184	0.026157277
B12	27.875	.715987102	12.66331125	0.001772125	0.00068321	0.001564729	0.000785346	0.007221765	0.012027176
B13	28.425	.3271111805	10.52469571	0.002747985	0.000776946	0.001584777	0.00050321	0.015172254	0.020785173
B14	30.2	.409396082	12.38081589	0.003120931	0.000290687	0.001703315	0.000836775	0.01652829	0.022479998
B15	27.975	.00143685599	10.31764107	0.00211228	0.000193012	0.000750602	0.001145771	0.004066245	0.008267911
B16	32.625	.8847944567	8.917867321	0.004879224	0.001563242	0.003954923	0.002459478	0.02972535	0.042582217
B17	30.625	.04474590248	8.905393214	0.00405686	0.001464366	0.000971551	0.000433125	0.012902288	0.01982819
B18	32.125	.2454288474	8.324514107	0.004014093	0.001024097	0.002398647	0.000467281	0.012330571	0.020234689
B19	32.25	.00058177055	11.75330482	0.004062375	0.001306952	0.003040376	0.000677774	0.023688011	0.032775488
B20	32.525	.795808045	10.07604304	0.002812599	0.000572938	0.001208413	0.000258933	0.012121138	0.01697402
B21	28.425	.2762734483	14.60917536	0.003544998	0.001568156	0.00261391	0.001106735	0.017818173	0.026651971
B22	29.5	.2084366097	15.29405339	0.004923521	0.000655771	0.002330931	0.001184612	0.017276033	0.026370869
B23	28.05	.9754244341	10.40844714	0.002847366	0.000771076	0.001770474	0.000657758	0.018985835	0.025032509
B24	30.8	.1037074883	10.94683518	0.002624943	0.000947219	0.000490868	0.000243463	0.014818279	0.019124771
B25	29.025	.9780120518	15.20716125	0.004633742	0.00186321	0.002422845	0.000613222	0.016629775	0.026162794

B26	30.025	.2670096963	12.45499464	0.004133144	0.001050177	0.002431245	0.001398164	0.015477225	0.024489954

Table 57: RPAV values of P.D. James' novels

Work	Richness	Gender	RA Power	Auditory	Gustatory	Haptic	Olfactory	Visual	Sensory
B1	30.4	0.940645	10.00232	0.002711	0.000563	0.001162	0.001051	0.007264	0.01275
B2	31.55	0.919922	10.80736	0.002661	0.000731	0.001719	0.000663	0.012255	0.018029
B3	34	0.889581	12.87791	0.001755	0.001393	0.002247	0.000259	0.014544	0.020197
B4	34.075	0.954264	15.18456	0.003247	0.001025	0.002996	0.001091	0.015896	0.024255
B5	33.375	0.853502	11.10274	0.002124	0.001027	0.001869	0.000174	0.014274	0.019467
B6	34.275	0.303183	11.87774	0.003538	0.000888	0.002602	0.000179	0.012304	0.019512
B7	30.05	0.9513	15.41543	0.004463	0.001192	0.002928	0.002557	0.015941	0.027081
B8	30.625	0.829785	10.91935	0.002079	0.000834	0.001406	0.000601	0.015989	0.020909
B9	34.325	0.978708	10.10973	0.003301	0.000322	0.001157	0.000612	0.015161	0.020554
B10	30.55	0.996497	13.41248	0.003375	0.001416	0.002557	0.001716	0.016954	0.026017
B11	31.525	0.99423	13.68775	0.003608	0.000752	0.001258	0.000254	0.013601	0.019472
B12	32.675	0.273332	11.98495	0.003458	0.000178	0.000888	0.00087	0.005123	0.010517
B13	32.7	0.995679	12.55905	0.002692	0.000833	0.001289	0.001319	0.01535	0.021483
B14	31.525	0.985426	12.88715	0.001869	0.000619	0.000807	0.000239	0.011505	0.01504
B15	26.375	0.047602	14.99576	0.002917	0.000479	0.00144	0.000517	0.011575	0.016928
B16	32.5	0.689278	11.40292	0.001975	0.000793	0.000481	0.00087	0.010119	0.014237
B17	32.575	0.888029	9.989731	0.002328	0.00022	0.000492	0.000163	0.006911	0.010114
B18	34.05	0.999632	14.26961	0.001656	0.000592	0.0023	0.000647	0.013249	0.018443
B19	28.9	0.847247	10.56264	0.001274	0.000403	0.000711	0	0.003552	0.00594

Table 58: List of Suicide attackers highlighting their different 'modus operandi' and event summaries

ID	Perpetrator	Date	wounded	killed	total	type of data	Location	Sex	Age	What
1	Joseph Stack	18-Feb-10	15	2	17	manifesto	Austin, Texas, USA	M	53	Aircraft-building
2	Charles J Bishop	1-May-02	0	0	0	suicide note	Tampa, Florida, USA	M	15	Aircraft-building
3	Jose Reyes	21-Oct-13	2	1	3	final letters	Sparks Middle School	M	12	gun - school
4	Karl Pierson	13-Dec-13	0	1	1	journal extracts	Arapahoe High School, USA	M	18	gun - school
5	Jiverly Wong	3-Apr-09	4	13	17	suicide note	American Civic Association, USA	M	41	gun - school
6	Pekka-Eric Auvinen	7-Nov-07	12	8	20	manifesto - 3 documents	Jokela High School, Finland	M	18	gun - school
7	Elliot Rodger	23-May-14	14	6	20	autobiography and retribution video translation	University of California, USA	M	22	gun - school
8	Duane Morrison	27-Sep-06	0	1	1	suicide note	Platte Canyon High School, USA	M	53	gun - school
9	Dorothy Dutiel	12-Feb-16	0	1	1	suicide note	Independence High School, Arizona, USA	F	15	gun - school
10	Gang Lu	1-Nov-91	1	5	6	Letters and statement	University of Iowa, USA	M	28	gun - school
11	Charles Whitman	1-Aug-66	32	16	48	Letters / suicide note	University of Texas, USA	M	25	gun - school
12	Michael Slobodian	28-May-75	13	2	15	suicide note	Centemmial Secondary School, Canada	M	16	gun - school
13	Adam Lanza	14-Dec-12	2	27	29	personal messages	Sandy Hook Elementary School, USA	M	20	gun - school
14	Myron May	20-Nov-14	3	0	3	final writing and letters	Florida State University, USA	M	31	gun - school
15	Wellington de Oliveira	7-Apr-11	12	12	24	suicide note	Escola Municipal Tasso da Silveira, Brazil	M	23	gun - school
16	Marc Lépine	6-Dec-89	14	14	28	suicide note	École Polytechnique, Canada	M	25	gun - school
17	Eric Harris	20-Apr-99	13	8	21	Journal	Columbine High School, USA	M	18	gun - school
18	Seung Hui Cho	16-Apr-07	17	32	49	school papers	Virginia Tech University, USA	M	23	gun - school
19	Robert Butler, Jr.	5-Jan-11	2	1	3	suicide note	Millard South High School, USA	M	17	gun - school
20	Bastian Bosse	20-Nov-06	37	0	37	Journal	Geschwister Scholl, Germany	M	18	gun - school

21	Dylan Klebold	20-Apr-99	110	5	15	Journal	Columbine High School, USA	M	18	gun - school
52	Robert A. Hawkins	5-Dec-07	2	10	12	Suicide Note	Westroads Mall, Nebraska, USA	M	19	gun - mall
53	Kyle Aaron Huff	25-Mar-06	2	6	8	Suicide Note	Capitol Hill, Seattle, Washington, USA	M	28	gun - party
54	Christopher Dorner	3-Feb-13	3	4	7	Manifesto	California, USA	M	33	gun
55	Mark Barton	29-Jul-99	13	12	25	Suicide Note	Georgia, USA	M	44	hammer / gun

External Data

Johnson Arc Sine Transformation of Richness for Shakespeare.
POS Analysis of Shakespeare data.

END OF THESIS