



**UNIVERSITY of the
WESTERN CAPE**



**INVESTIGATING A PROPOSED STANDARDISED METHOD FOR
AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS**

BY

ZAHN-MARI KOTZÉ

THE UNIVERSITY OF THE WESTERN CAPE

**Thesis submitted in fulfilment of the requirements for the degree of
MASTER OF ARTS**

in the

Department of Linguistics

Date of submission: 21 April 2023

Supervisor: Professor Monwabisi K. Ralarala

Co-supervisor: Doctor Annelise de Vries

ABSTRACT

Authorship attribution is a method of revealing the obscure or unknown individuals who may have played a part in the creation of texts (Kotze, 2007). The purpose of authorship attribution is both to test claims of authorship and to detect secret or anonymous authors. Yet, to the best of our knowledge, there is no standardised method for authorship attribution. The literature reveals a need for such a method to be devised and ratified by the courts so that forensic linguists may act as expert witnesses. Such a method needs to be both quantitative and qualitative in nature. The dearth of such a method led to a mixed-method investigation into existing frames and methods of authorship attribution, with a view to proposing a method of enquiry that would uncover the unseen authors or contributors to texts. A detailed and systematic literature review led to the identification of writing-style features and a classification technique for the proposed method.

The research therefore set out to discover whether authorship attribution could include a consideration of T-units and cohesion markers, in addition to Chaski's (2007: 133–146) existing language indicators of authorship, which are (i) end-of-sentence punctuation, (ii) internal structure of sentences, and (iii) average sentence length. Ultimately, the investigation sought to uncover whether the additional two markers could be used for authorship profiling among selected first language (L1) and second language (L2) English speakers. These markers constituted the specific writing-style features of the various authors and were manually tagged in the chosen texts. The classification technique was assisted by the software, *WordSmith Tools*. The investigation analysed the linguistic evidence of eight L1/L2 English texts – four L2 texts and four L1 texts – to test whether the methodology actually worked. The markers were identified in the texts, and graphs and tables were calculated to present the quantitative data. From this data, linguistic deductions were made about the authors.

The study found that a method for hidden authorship attribution is possible, but its success depends on the combination of writing-style features selected and the nature of the text analysed. The study found that average sentence length, average T-unit length and cohesion markers are good indicators of authorship, but that a more reliable classification technique is still needed. However, the current method may be used successfully in the analysis of texts of

various lengths found online, as long as there is more than one text by the same author to serve as a basis for identifying these features.

Key words: Authorship attribution, Standardised method, Writing-style feature, Classification technique, End-of-sentence punctuation, Internal structures of sentences, Average sentence length, T-units, Cohesion markers, First language speakers (L1), Second language speakers (L2), Linguistic evidence



DECLARATION

Name: ZAHN-MARI KOTZÉ

Student number: 4255309

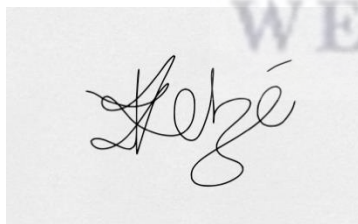
Degree: MASTER OF ARTS

*INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS*

I declare that the above thesis is my own work and that all the sources I have used or quoted have been indicated and acknowledged by the means of complete references.

I further declare that I submitted the thesis to originality checking software and that it falls within the accepted requirements of originality.

I further declare that I have not previously submitted this work, or part of it, for examination at the University of the Western Cape for another qualification or at any other higher education institution.



21 April 2023

SIGNATURE

DATE

ACKNOWLEDGEMENTS

First, I would like to thank my Lord and Saviour, Jesus Christ, for none of this would have been possible without His plan and guidance. During hard times, He is the first I consult and look to for strength, hope, joy, perseverance and peace.

I sincerely thank my supervisors, Prof Monwabisi Ralarala and Dr Annelise de Vries, for their continued support, guidance and knowledge. I would not have completed this thesis so quickly if it were not for their belief in my abilities and encouraging words and actions. I am a blessed woman having your support and knowledge at my disposal.

To my parents, Corlie and Stephan; thank you so much for always being there to listen and to lend a shoulder to cry on when I felt overwhelmed. Thank you for allowing me to miss days at work to focus on my studies or attend functions and writing retreats. You sacrificed a great deal to give me a future. By being such hardworking people yourselves, you shaped me into the hardworking woman I am today. I am proud of who I am today thanks to you. Good job!

To my aunt and uncle (my second set of parents), Prof Derick Blaauw and Prof Anmar Pretorius; thank you for being examples that I will try to emulate. You have helped and supported me all my life, whether through moral support, driving me around or supplying comedic relief. You have helped me immensely with the rounding-off of this thesis, using your knowledge of academic writing, statistics, numbers, referencing, etc. I will always be thankful to have you in my life, leading the way.

A heartfelt thank you, too, to all my other family members and friends who are always there with an encouraging chat or a coffee (or a glass of wine). Thank you for keeping me in your prayers and believing in me even when I struggled to do so myself.

A special thanks to Prof Wannie Carstens, who was the inspiration for combining text linguistics and forensic linguistics in my research; and to Dr Karien van den Berg, who introduced me to the field of forensic linguistics and sparked the idea for this research. Also a special thanks to Professor Hilton Hubbard, who generously gave me the L2 texts for analysis.

This work was supported by the National Institute for the Humanities and Social Sciences (NIHSS). Opinions, findings, conclusions and recommendations expressed in this publication are that of the author; the NIHSS has no liability in this regard.



TABLE OF CONTENTS

KEY WORDS	ii
ABSTRACT	II
DECLARATION	iv
ACKNOWLEDGEMENTS	v
CHAPTER 1: INTRODUCTION	1
1.1 Context of the study	Error! Bookmark not defined.
1.2 Authorship analysis defined.....	3
1.3 Background to the research.....	4
1.4 Problem statement	5
1.5 Research questions.....	6
1.6 Objectives of the study	7
1.7 Significance of the study.....	7
1.8 Research methodology.....	8
1.8.1 Data.....	8
1.8.2 Method	8
1.8.3 Ethical clearance	9
1.9 Chapter breakdown.....	9
1.10 Conclusion.....	10
CHAPTER 2: LITERATURE REVIEW	
2.1 Introduction.....	Error! Bookmark not defined.
2.2 Forensic linguistics: Defining the field and sub-categories	11
2.2.1 The origin of the discipline.....	15
2.2.2 Theoretical linguistics vs. applied linguistics	21
2.3 Authorship analysis in South Africa: An emerging sub-category.....	23
2.4 Authorship attribution/identification	24
2.4.1 Where it started	24
2.4.2 Research in the field of author identification	25
2.5 Author identification and the court.....	29
2.5.1 The forensic linguist's role in court cases	29
2.5.2 Forensic-linguistic evidence in courts: A brief overview	30
2.6 Stylometry	36
2.6.1 Adversarial stylometry	40
2.6.2 Shorter texts and stylometry.....	43
2.7 Idiolect	44
2.7.1 What is generic language use?	44
2.7.2 What is idiolect?.....	45
2.7.3 The problems surrounding idiolect.....	45
2.7.4 Theories about the presence of idiolect	49
2.8 Writing-style features used in stylometry	52
2.8.1 Lexical features	52
2.8.2 Syntactic features	53
2.8.3 Structural features	54
2.8.4 Content-specific features.....	54
2.9 Computer-based methods/Classification techniques used in stylometry	55
2.9.1 CUSUM technique (CUMulative SUM).....	56
2.9.2 N-gram analysis	56
2.10 Writing-style features identified for this study	59
2.10.1 Variables identified by Carole Chaski.....	59
2.10.2 T-units	60
2.10.3 Cohesion markers.....	61
2.11 Conclusion.....	Error! Bookmark not defined.
CHAPTER 3: METHODOLOGY AND ANALYSIS OF DATA	70
3.1 Introduction.....	70
3.2 Research approach	70
3.2.1 A corpus-based approach: A brief overview	70

3.2.2	Types of research done using a corpus-based method.....	72
3.3	Research questions.....	73
3.4	Research design	73
3.4.1	Nature of reasoning	74
3.4.2	Type of data	74
3.5	Data analysis.....	75
3.5.1	Phase 1: Data collection.....	77
3.5.2	Phase 2: Feature extraction (also known as tagging).....	77
3.5.3	Phase 3: Method generation	78
3.5.4	Phase 4: Authorship identification/attribution.....	78
CHAPTER 4: FINDINGS AND DISCUSSION		80
4.1	Introduction.....	80
4.2	Data presentation	80
4.2.1	First language English speakers' texts.....	81
4.2.2	Second language English speakers' texts	88
4.3	Feature extraction	98
4.3.1	Average sentence length.....	99
4.3.2	Internal structure of sentences.....	99
4.3.3	End-of-sentence punctuation.....	99
4.3.4	Average T-unit length	99
4.3.5	Average number of T-units	99
4.3.6	Reference	100
4.3.7	Substitution	100
4.3.8	Ellipsis.....	100
4.3.9	Conjunction.....	100
4.3.10	Lexical cohesion.....	101
4.4	Findings	101
4.5	Feature extraction from texts: Some examples	104
4.6	Discussion of findings	106
4.6.1	Average sentence length.....	107
4.6.2	Internal structures of sentences	107
4.6.3	End-of-sentence punctuation.....	107
4.6.4	Average T-unit length	108
4.6.5	Average number of T-units per sentence	108
4.6.6	Reference	108
4.6.7	Substitution, ellipsis, conjunction and lexical cohesion	109
4.7	Statistical significance of the study.....	111
4.8	Successful writing-style features	112
CHAPTER 5: CONCLUSION AND RECOMMENDATIONS		115
5.1	Answers to the research questions.....	115
5.2	Summary of the research findings.....	116
5.3	Limitations of the study.....	117
5.4	Recommendations for further study.....	118
5.5	Conclusion.....	119
REFERENCES.....		ERROR! BOOKMARK NOT DEFINED.
Appendix 1.....		1366
Appendix 2.....		137
Appendix 3.....		138
Appendix 4.....		139
Appendix 5.....		142
Appendix 6.....		146
Appendix 7.....		150
Appendix 8.....		154

LIST OF TABLES

Table 2.1: Categories within forensic linguistics.....	15
Table 4.1: L1 Dataset size.....	81
Table 4.2: L2 Dataset size.....	Error! Bookmark not defined.
Table 4.3: Feature extraction colour coding	Error! Bookmark not defined.
Table 4.4: Results of L1 texts' analysis	Error! Bookmark not defined.
Table 4.5: Results of L2 texts' analysis	Error! Bookmark not defined.
Table 4.6: Some examples of feature extraction of L1 texts ...	Error! Bookmark not defined.
Table 4.7: Some examples of feature extraction of L2 texts ...	Error! Bookmark not defined.
Table 4.8: Test of significance between means of L1 and L2 .	Error! Bookmark not defined.

LIST OF FIGURES

Figure 3.1: Conceptual framework for authorship attribution.....	76
Figure 4.1: Average sentence length.....	107
Figure 4.2: Average T-unit length	108
Figure 4.3: L1 cohesion markers	110
Figure 4.4: L2 cohesion markers	110
Figure 4.5: Comparison between L1 and L2 average T-unit length relative to average sentence length.....	113

CHAPTER 1: INTRODUCTION

1.1 Context of the study

Over the past decade, it has become clear that linguists have the potential to be of great assistance to the legal system and courts, especially in Common Law countries, which increasingly rely on language experts for assistance in cases involving authorship identification, voice identification, plagiarism, statement analysis and legal interpretation and translation. Cases have arisen in which there is linguistic proof of a statement, but several suspects, or none. In order to identify the author in such cases, a standardised authorship attribution method would be useful, along with a standardised way of determining whether the text was authored by a first language speaker (L1) or second language speaker (L2); in other words, a standard authorship profiling method.

The usual starting point for authorship analysis is the assumption that every native speaker has an idiolect and uses language in a unique way. Forensic linguists assume that idiolect will reveal itself through unusual and idiosyncratic speech and writing choices (Coulthard & Johnson, 2007). Idiolect, linguistic profiling and authorship attribution rely on the Theory of Markedness (Olsson, 2008) to identify the distinctive linguistic characteristics of a speaker/author. These can be used to characterise the text in an attempt to identify the speaker/author.

Perhaps the most famous case in the history of forensic linguistics is that of Ted Kaczynski, the Unabomber. Coulthard, Johnson and Wright (2016) provide the best description of the case's events. Between 1978 and 1995, an American who identified himself as 'FC' left bombs in the mail on average once each year. In 1995, six national newspapers received a 35,000-word manuscript titled 'Industrial Society and its Future' from an individual claiming to be the Unabomber. Three months later, a man called the FBI claiming that the document sounded like it was written by his brother. He used the phrase 'cool-headed logician' as an example of a phrase used in the document that was typical of his brother's idiolect. This led to the capture of the Unabomber. However, the difficulty with the use of linguistics in investigations is that investigative work is complex and diverse; each case may necessitate the development of a unique methodological approach.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

The finest known use of authorship attribution in the setting of forensic linguistics in South Africa was recounted by Prof Hilton Hubbard, who testified in a 1989 extortion case (Hubbard, 1994; 1995). This case was based on a study of a series of 10 extortion letters received by the Johannesburg branch of a national supermarket chain, which contained threats to poison food on the company's shelves unless R1.5 million (about R6 million today) was paid to the author. The defendant in the case was a native Polish speaker with a less-than-perfect command of English. In support of the prosecution's argument, error analysis was conducted on two sets of texts: the letters received by the company and a total of seven essays of varying lengths that an expert witness persuaded the defendant, and others, to write. While it was evident that error analysis¹ alone would not be sufficient to win a conviction (SCSA, 1989), Hubbard stated that the error patterns in the accused's writings had to be quantitatively compared with those of a person with a comparable background and English competence. According to Hubbard, such a comparison would have the highest chance of producing a correct conclusion. However, in the end, the texts were deemed too brief and the sample size too small to withstand the rigour of statistical significance testing.

A more recent case involving authorship attribution in South Africa, *Zulu v. Mathe* (2021), raised the question of whether linguistic experts have or use the proper tools for authorship attribution. The case came about after the passing of the late King Goodwill Zwelithini Zulu, and the run-up to the coronation of Prince Misuzulu KaZwelithini. The Zulu princesses Ntandoyenkosi Zulu and Ntombizosuthu Zulu-Duma filed an interdict to halt Prince Misuzulu's coronation. They asserted that the deceased King Goodwill Zwelithini's will was forged. Although the princesses attempted to halt the coronation, they did not contest the royal family's election of Prince Misuzulu as Zulu king. However, they successfully halted the execution of Zwelithini's will pending a trial to determine its legality. The interdict lasted 15 days before expiring.

Two witnesses testified regarding the will, but their opinions were contradictory. The authenticity and validity of the deceased king's will were in question. Mr Yossi Vissoker and Mr Cecil Greenfield were instructed to compare the purported signatures of the deceased king

¹ Error analysis is a method used to document the errors that appear in an author's language, to determine whether those errors are systematic, and (if possible) to explain what caused them.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

in the will dated 29 November 2016 with the deceased king's known signatures, and to render an opinion on their authenticity. The two experts made different findings and drew different conclusions; one expert asserted that the signatures were forged, while the other asserted that they belonged to the deceased king. The handwriting report prepared by Vissoker claimed that grave discrepancies in the signatures indicated that they were not signed by the same hand, while Greenfield asserted that variations in signatures were not uncommon. Variations in strokes, forms and characteristics of letters result from the inability of the human hand to write with mechanical precision. Common cause was the fact that the signatures differed in some way; however, Mr Vissoker attributed this difference to forgery, while Mr. Greenfield attributed it to natural variation. This created a sharp factual dispute for the case, as well as for forensic linguists. This case brought into question the reliability of certain methods and how objective such methods are. Tools and methods are needed that are accepted by everyone.

Forensic linguistics is a nascent field of study internationally, but especially in Africa and South Africa. As a result, there is still a great deal of research to be conducted on the creation of innovative and reliable approaches and methods in this field.

1.2 Authorship analysis defined

Dr Andrea Nini (n.d.) defines authorship analysis as ‘... the application of linguistic methods to shed light on the authorship of a questioned text’. For instance, it can be used to indicate the most likely author of a text from a sample of suspects (authorship attribution/identification) or the most likely demographic details of an anonymous author (i.e., authorship profiling). These techniques are commonly adopted in forensic linguistics to solve cases of disputed authorship, including cases of threatening, abusive or generally malicious texts. Authorship profiling is defined by Nini (2015) as the task of determining information about the background of the author of an anonymous text based on language use in the text. Authorship profiling is done in an effort to determine age, gender, language proficiency and education level of the author of text or texts.

In this dissertation, authorship attribution is distinguished from authorship identification, based on the argument made by Olsson (2008, p. 45):

My reason for disliking this term [‘authorship identification’] is that, again, if we have two possible authors as candidates, we are not really undertaking an identification exercise. To say

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

that our writer is more likely to be x or more likely to be y is not the same as saying that our writer is x or y . We are not actually identifying x or y . We are proposing a greater likelihood of x than y , or the other way around. If we were absolutely certain that x was the writer, for example, then we could say we are identifying him or her, but it is a basic premise of any scientific method that results are not given as certainties, but as probabilities or classifications. It is very rare in authorship work that we can *identify* a writer. Also, note that identification here suggests being able to pick out one individual author from all other authors.

For the purpose of this dissertation, authorship attribution is the preferred term. As Olsson states above, attribution is more possible than identification, although, of course, the hope is always to identify an author with 100% certainty. However, more often than not, it is more feasible to attribute a text to the most likely author based on the linguistic analysis conducted. This dissertation in fact makes use of both terms, but that is only because most sources use the term ‘authorship identification’.

1.3 Background to the research

Kotzé (2010) states that authorship identification and authorship attribution are broad and interdisciplinary subjects of study, having applications in religion, literature, education, national and commercial intelligence, and, of course, the practice of civil and criminal law. Until recently, the forensic application of author identification techniques was largely restricted to handwriting specialists brought in as expert witnesses; methods applied in non-linguistic forensic fields were largely restricted to disputes over the authorship of literary, theological, philosophical and political documents (Kotzé, 2010). However, the focus of forensic authorship identification has gradually and decisively shifted from procedures such as analysis of handwriting, graphical features and writing materials to the linguistic content of legally significant documents such as blackmail letters, confessions, wills, suicide letters and plagiarised writings (Kotzé, 2010). In this use, ‘linguistic content’ is identified and analysed by a skilled linguist and is predominantly qualitative in nature. Some believe that the linguist’s job consisted solely of comparing texts based on a stylistic analysis. Stylistic analysis, according to Guillén-Nieto, et al. (2008), is an approach to authorship identification in literary contexts, predicated on the idea that it is possible to identify, describe and quantify an author’s individual style or idiolect through careful linguistic observation and analysis of their unique linguistic choices. In the context of litigation, forensic text analysis uses stylistics to reach a conclusion and form an opinion regarding the authorship of a contested text (McMenamin, 2002). Even though linguistic analysis has been qualitative in nature in the past, as the sciences have progressed, the quantitative component is moving to the fore.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

The quantitative method of text analysis, or stylometry, has expanded dramatically with the arrival of the technological age (Kotzé, 2010). Recent changes in the criteria for reliability and evidence, particularly in the United States (*Daubert v. Merrell Dow Pharmaceuticals* (92–102, 509 U.S. 579, 1993)) have emphasised the scientific method traditionally used in the natural sciences, and the need to provide quantitative or measurable probability with regard to the admissibility of linguistic evidence in court (Nieto, et al, 2008).

Daubert v. Merrell Dow Pharmaceuticals (92–102, 509 U.S. 579, 1993) gave rise to the Daubert standard on expert witnesses. The Daubert standard is aimed at determining whether the evidence of an expert makes use of a scientifically valid methodology that can be applied to the relevant facts (Nieto, et al, 2008). Since the methods of enquiry in the humanities and social sciences are unavoidably more relative than those in the natural sciences, this poses a threat to the linguist as an expert witness in the legal setting. However, numerous linguists concur that the current emphasis on quantification is significant for two reasons: (a) it satisfies current methodological requirements for the study of linguistic variation (hypothesis testing and verification), and (b) it satisfies external judicial requirements for expert testimony (McMenamin, 2002).

1.4 Problem statement

According to Rudman (1998), over 1,000 stylometric writing-style features ('style markers') have been proposed. However, no set of significant writing-style features or style markers have been identified as uniquely discriminatory. Furthermore, some proposed writing-style features may not be valid discriminators; for example, prescriptive grammar errors and profanities. These are not generally considered to be idiosyncratic, i.e. unique or distinctive (Zheng, et al, 2006). Just as there is a range of available writing-style features, there are many different classifications of techniques for authorship identification/attribution. These include statistical approaches (e.g., the CUSUM, Thisted and Efron tests), neural network approaches (e.g., the use of radial basis functions, feedforward neural networks, cascade correlations), genetic algorithms and Markov chains (Zheng, et al. 2006).² However, there appears to be no consensus on a correct methodology, with many of these techniques suffering from problems such as

²These are all examples of statistical/quantitative methods or software used for authorship analysis, also known as classification techniques.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

questionable analysis methods, inconsistencies within the works of one author and/or failed replication.

This study investigates what writing-style features and classification techniques have been used in the past, and why each was deemed successful or unsuccessful. This review of methodologies already in existence assisted in the identification of features and techniques which may be used in the proposed standardised method. The proposed methods also introduce new features and techniques which may not have been well explored before.

In addition, the research investigates whether it is possible to differentiate between L1 and L2 authors of linguistic evidence used in court using a standardised approach to authorship attribution. The question was whether or not the language features would be definitive enough to permit profiling of L1 and L2 English speakers as authors, and whether qualitative and quantitative analysis could be combined to produce a comprehensive methodology applicable in all cases.

South Africa is a linguistically diverse country with eleven official languages, which necessitates that a methodology be devised that can reliably and consistently distinguish between L1 and L2 English speakers. The majority of the population do not speak English as a first language; in fact, only 8.1 percent of the population are first language English speakers. As of 2018, the languages most commonly spoken by individuals in South African households were isiZulu, at 25.3 percent, isiXhosa, at 14.8 percent and Afrikaans, at 12.2 percent. While English is only the sixth-most common language spoken in South African households at 8.1 percent, it is the second-most prevalent language spoken outside of homes, at 16.6 percent (Galal, 2022).

1.5 Research questions

The following questions arise from the literature. They indicate the purpose of this research:

- i. To what extent is a standardised method for authorship identification/attribution possible, making use of a certain combination of writing-style features?
- ii. What are the available writing style features, and to what extent are they effective for authorship identification/attribution of L1 and L2 English texts?
- iii. Which classification techniques are effective for authorship identification/attribution of L1 and L2 English texts?

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

- iv. Can the proposed standardised method be deployed on readily available online texts?

1.6 Objectives of the study

Drawing from the above discussion, the objectives of this study are the following:

- i. to investigate to what extent a standardised method for authorship attribution is possible;
- ii. to identify the available writing style features, and the extent of their effectiveness for authorship identification/attribution of L1 and L2 English texts;
- iii. to identify the classification techniques and the extent of their effectiveness for authorship identification/attribution of L1 and L2 English texts; and
- iv. to contribute to the study of forensic linguistics with particular reference to authorship attribution.

The objectives were carefully chosen to assist in the creation of a possible standardised method for authorship attribution of L1/2 English texts. If the proposed method proves successful, it will make a noteworthy contribution to authorship attribution and authorship profiling in the field of forensic linguistics, and also to the rich multilingual landscape of South Africa. Ultimately, the aim is to create a method that is acceptable to the courts and satisfies the external requirements for expert evidence as imposed by the judiciary.

1.7 Significance of the study

The research investigates whether it is possible to differentiate between L1 and L2 authors of linguistic evidence using a reliable approach to authorship attribution that employs a specified set of features. In a country as linguistically diverse as South Africa, such an approach is of crucial importance. Since South Africa has eleven official languages and only 8.1 percent of the population speaks English as a first language (Galal, 2022), anonymous content used in court is typically written by non-native English speakers. A method for differentiating between L1 and L2 English speakers would be extremely useful. Given that forensic linguistics is a rapidly expanding field in South Africa, this research is therefore both significant and innovative. The research aims to change the fact that, to the best of our knowledge, there is currently no approved or standard method for authorship attribution.

1.8 Research methodology

1.8.1 Data

The study used readily available data obtained from the internet and books. Texts were chosen based on their relevance and availability, and purposefully sampled as examples of L1 and L2 English texts in light of the objectives of the study. Linguistic evidence made available by Prof Hilton Hubbard (Hubbard, 1994; 1995) from the 1989 extortion case detailed earlier, was used. Six L2 texts were received by Prof Hubbard, but only four were used, as this was the number of L1 texts available, and the author wished to have balanced numbers of both kinds of text. The four L1 texts were found in the appendices of various forensic linguistics handbooks. These texts are all letters threatening bomb explosions: 'The Army of God' letter, the 'Lampley Hollow' letter, a letter by Luke Jon Helder (a pipe bomber) and the letter sent by Theodore Kaczynski (the Unabomber, as discussed under 1.1, Context of the study). These texts' origins are thoroughly explained in Chapter Four. Multiple texts by the same author would have been ideal, as these would have revealed when and how often an author used their chosen writing-style features.

1.8.2 Method

Using the conceptual framework proposed by Zheng, Li, Cheng and Huang (2006) (see Section 3.5) as a foundation for the methodology, the following steps were undertaken:

- The three language elements (end-of-sentence punctuation, internal structure of sentences and average sentence length) of Chaski's programme, ALIAS³, were identified in the texts.
- The T-units were identified.
- The cohesion markers were identified.
- These features were then quantified in order to determine how frequently each occurred and which features were most prominent in the writing of individual writers.

The steps were employed for all eight L1 and L2 English texts chosen, a process which yielded answers to the research questions. The features of the individual texts were compared to

³ Chaski (2007) developed the software, ALIAS Technologies, hereinafter referred to as ALIAS.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

determine whether L1 and L2 English speakers make fundamentally different choices with regard to the writing-style features identified in Section 2.9.

1.8.3 Ethical clearance

Ethical clearance was not required for this research since the study relied on secondary data, as stated in Section 1.8.1.

1.9 Chapter breakdown

Chapter One presents context of the study, states the research questions and objectives, and introduces the concept of authorship attribution in the field of forensic linguistics.

Chapter Two discusses the broad field of forensic linguistics, situating this study within the discipline. The origin of authorship attribution is discussed to yield a clear view of what has been achieved in this sub-category from its inception. The chapter highlights the position of forensic linguistics in South Africa, showing that very little research exists on authorship attribution on the African continent, to the best of our knowledge. This part of the chapter also reviews the literature on authorship analysis in court, along with the literature on the forensic linguist's role in court as an expert witness. The review reveals that there is still a need for an authorship attribution method that is accepted by the courts. The chapter then moves on to the area known as 'stylometry', finding that it consists of two main processes: the identification of characteristics and the statistical processing of these features using a classification method (Barry & Luna, 2012). Both aspects are necessary because courts want expert witnesses to report on their findings based on quantitative proof. The chapter discusses the role of idiolects, and the role they play in authorship analysis. As explained in Section 2.8, an idiolect is a person's particular, distinctive use of language, created subconsciously. Individuals are typically unaware of the idiolectic terms they employ when speaking or writing, and the idiolectic methods they typically use. The various definitions of idiolect that exist are based on the core concepts of 'individuality' and 'uniqueness'. The chapter identifies existing writing-style features and classification techniques used for authorship attribution, and proposes new features that may be used in the identification of idiolect, and which could possibly lead to a successful method for authorship attribution.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

Chapter Three discusses the methodology used in this research to examine eight texts of interest. The steps used in this examination were (i) data collection, (ii) feature extraction, (iii) method generation and (iv) authorship identification/attribution. A mixed-method approach was applied to eight L1/L2 texts: four L1 texts obtained from books by Olsson (2008) and Gales (2010); and four L2 texts received from Professor Hilton Hubbard, from the 1989 extortion case he worked on.

Chapter Four presents the findings and discusses them in light of the literature. The writing-style features that are tagged and quantified using *WordSmith Tools*⁴ are presented in separate tables for the L1 authors and the L2 authors, so that the texts may be compared and deductions made.

Chapter Five concludes the study and makes recommendations for future research.

1.10 Conclusion

Chapter One has established the context of the research, giving some idea of the value of forensic linguistics and explaining the need for a standardised methodology for authorship attribution of L1 and L2 English users.

The following chapter discusses the field of forensic linguistics in light of the literature, indicating where authorship analysis is situated within the discipline. It also gives a detailed account of the field of stylometrics for authorship attribution.

⁴ Scott (2021) is the developer of *WordSmith Tools*, hereinafter referred to as *WordSmith* or *WordSmith Tools*.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

Forensic linguistics is a rapidly expanding sub-field of applied linguistics and a well-established field of linguistic enquiry. In its broadest definition, forensic linguistics refers to the study of language in legal and investigative contexts. The field can be further subdivided into three overlapping areas: (i) investigative forensic linguistics, (ii) the study of the written language of law, and (iii) the study of communication in the legal process (Coulthard & Johnson, 2010; Coulthard et al., 2011; Perkins & Grant, 2013). It is in the subdivision of 'investigative forensic linguistics' that this research is situated. Investigative forensic linguistics involves comparative authorship analysis, sociolinguistic profiling, a determination of disputed meaning, native language identification, and more (Tkacukova, 2019). Authorship identification based on authorship analysis is a method of revealing obscure or unknown authors, secret complicity in the creation of texts, or an author's indirect involvement in their creation (Kotzé, 2007).

2.2 Forensic linguistics: Defining the field and sub-categories

This chapter discusses forensic linguistics as a subject of study. The origins of the field of study are examined, with a particular emphasis on author identification, the focus of the current research. Stylometry, a technique frequently employed in forensic linguistics, is examined and discussed, in particular the more recent subdivision of stylometry known as adversarial stylometry. This branch of stylometry has risen to prominence as issues to do with online privacy have assumed centre stage. Adversarial stylometry examines author identification from the perspective of deliberate attempts to hide identity, a common practice in the online environment. The term idiolect is also discussed, and efforts are made to ascertain whether it is feasible to speak of an individual's idiolect and whether such idiolect can be identified in the texts of individuals.

First, it is essential to establish the place of forensic linguistics within the area of forensic science, given that linguistic analysis and its results may be used in court cases to identify suspects.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

The term 'forensic' is associated with terms such as 'judicial' and 'legitimate' and has historically been connected with science. For many years, forensic science has been used to catch criminal suspects and establish their guilt or innocence. Among the many components of forensic science are DNA analysis, fingerprint analysis and bloodstain analysis. Handwriting analysis and signature analysis, often used in forensic linguistics, are also considered part of forensic science (Jackson & Jackson, 2004).

Forensic linguistics is usually limited to linguistic features such as language use, voice characteristics, and the meanings of words within a given context. Although forensic linguistics is related to conventional types of analysis such as handwriting analysis, it focuses on other aspects of text, since the distinguishing characteristics of written texts are absent from modern electronic texts. For instance, all letters are typed uniformly, with the exception of variances caused by bolding or italicising letters or words.

According to Coulthard and Johnson (2007), the term forensic linguistics first appeared in 1949 in F.A. Philbrick's book *Language and the law: The semantics of forensic English*, which dealt with judicial English. Philbrick (1949, p. vi), used the term 'forensic' to refer to 'the English employed by attorneys and judges in legal proceedings'. However, the term was not immediately applied to language or linguistics. Olsson (2004) and Blackwell (2012) note that the term 'forensic linguistics' was first used with its current meaning in 1968 with the publication of Svartvik's examination of the Timothy Evans statements, titled *The Evans statements: A case for forensic linguistics*. Researchers agree that the term 'forensic linguistics' was still not widely recognised until the 1990s, when forensic linguistics became a recognised field of study. Initially, however, the methodologies used by linguists to conduct some of the earliest forensic linguistics studies could not be categorised under a single academic subject, as these analyses frequently made use of entirely new methods rather than applying an existing method (Blackwell, 2012).

Forensic linguistics is a subfield of applied linguistics in which linguistic knowledge and methods are used to generate a range of texts (spoken and written) for legal purposes. Olsson's (2008, p. 3) definition of forensic linguistics demonstrates forensic linguistics' close relationship with the judicial system:

Forensic Linguistics is the interface between language, crime and law, where law includes law enforcement, judicial matters, legislation disputes or proceedings in law, and even disputes

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

which only potentially involve some infraction of the law or some necessity to seek a legal remedy.

Leonard (2006, p. 65) condensed the preceding definition into a single sentence: 'Forensic linguistics applies the well-established field of linguistics to legal language data.' This definition highlights one of the most fundamental aspects of forensic linguistics, namely that it is an applied science. Since legal difficulties may arise in any profession or sector, it is important to keep in mind that forensic linguists may be involved in many fields. This means that the forensic linguist must be well versed in linguistics and comprehend the structure and vocabulary, dialects, sociolects and registers of a variety of language users. The forensic linguist also requires a basic background or understanding of the particular field or profession in which he/she performs investigations.

Thus forensic linguists need not only legal expertise, but also a measure of knowledge of the particular field in which they work. In addition, forensic linguists must be well versed in various elements of the language at play. McMenemy (2002) asserted that the linguists who pioneered forensic linguistics frequently proclaimed that they performed linguistics that happened to occur within a forensic setting, and that the forensic linguist therefore primarily needed a solid foundation in linguistics. According to Leonard (2006), forensic linguists bolster judicial cases by using rigorous, scientifically acknowledged principles of linguistic analysis as part of legal evidence. McMenemy and Leonard do not downplay the significance of legal knowledge, but emphasise the significance of knowledge in the field of linguistics and applied linguistics, and the theories and methodologies used to analyse language.

Today, various categories have been established within forensic linguistics. Table 2.1 gives an overview of these categories and an explanation of each.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

Table 2.1: Categories within forensic linguistics

Category	Description and focus
1. Language use related to legal implications	The topic of enquiry is the language employed in legal papers and legal proceedings. The forensic linguist explores the accessibility of legal language and focuses on the semantic side of language. Establishing the meaning of words within a certain (legal) context is crucial. Other aspects of legal and law-related language are also examined.
1.1 Court interpreting and translation	The focus is on courtroom interpretation and translation. This includes translating statements and other papers and analysing the evidence. Here, accuracy in interpreting and translation, the duty of the interpreter, the licensing of interpreters, ‘control’ over the person for whom interpreting is conducted, etc., are the primary concerns.
1.2 Language use and discourse in court	In this aspect, factors such as the relationship between legal entities and suspects in court, along and the language they employ (such as intimidating or manipulative language) are examined. Power, partisanship, and culture are investigated.
1.3 Transcription of verbal statements	When verbal statements are transcribed for the court, the completeness of these statements and the risk of bias in the transcript are evaluated.
1.4 Language rights (this category’s investigations go wider than the courtroom)	This aspect examines, among other things, the language rights and language use of minority languages – how these groups are dominated by other languages or dialects of the same language, and how bureaucratic language use oppresses individuals.
2. The examination of forensic texts (both spoken and written texts)	The implementation of the law includes forensic research and investigations conducted on texts outside of court. Forensic linguists are frequently contacted by the police to analyse the relationship between certain texts and crimes. These enquiries take several forms.
2.1 Author identification	Author identification is the process of analysing a text to determine its likely author, or the authors of many written works. Texts may comprise threats, blackmail letters, defamation messages, suicide notes and ransom demands, among others.
2.2 Profile composition	This branch of study is an alternate kind of author identification in which a forensic linguist creates a profile of a possibly suspect author based on linguistic evidence in a specific text. It is possible to determine the author’s age and gender based on linguistic indicators such as word usage, punctuation and the proportion of capital to lowercase characters.
2.3 Identification of plagiarism	Identification of plagiarism is particularly important in academic settings. Today, computer algorithms can detect instances of plagiarism, but in some instances, the expertise of a specialist is still required.
2.4 Speaker identification (also known as forensic phonetics)	This sub-category pertains specifically to the identification of a message’s speaker based on acoustic properties and sound characteristics of a voice. These messages may include telephone messages, emergency calls, recordings and telephone threats.
2.5 Forensic dialectology	This aspect examines the linguistic background of asylum seekers. It is possible to determine whether an asylum seeker truly belongs to a

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

	particular language group or cultural group through the use of tests of pronunciation administered to asylum seekers.
--	---

Adapted from Olsson (2004, p. 4-5)

Further topics of research, such as trademark ownership investigations, are difficult to fit under the aforementioned categories, but are considered forensic linguistic fields of study. Coulthard and Johnson (2007) state there are five broad areas of interest under which the knowledge of linguists who serve as courtroom witnesses might be categorised. According to Coulthard and Johnson (2007), trademark research falls under the first category, *Morphological meaning and phonetic similarity*. The remaining four classes are:

- *Syntactic complexity in a letter*. The intricacy of syntax may also apply to lengthy texts. In such situations, the linguist must assess if the intricacy of the sentence structure in a letter or document hinders the reader's comprehension of the information.
- *Lexico-grammatical ambiguity*. This category pertains to ambiguities encountered in texts; hence, the real meaning of certain words or phrases in specific situations/contexts must be determined.
- *Lexical meaning*. In this category of enquiry, it is only necessary to establish the meaning of specific words. Context, culture and the origin of words are considered.
- *Pragmatic meaning*. This category focuses mostly on the manner in which the message is conveyed. In this context, elements such as shared knowledge between participants in a communication act are relevant. In this area, one determines how realistic certain alleged statements are and what information is required for a legitimate confession of guilt.

2.2.1 The origin of the discipline

The origin of forensic linguistics is a controversial question. During the 1960s, 1970s and 1980s, linguists' knowledge was used in court cases where linguistics and law overlapped, mostly in the United States and Canada. However, such instances were few, and the approaches employed had not been thoroughly evaluated to determine their validity and dependability (Turell, 2008).

The Evans statements: A case for forensic linguistics by Svartvik is generally regarded as the earliest documented application of forensic linguistics (Coulthard, 2004; Olsson, 2004; Turell, 2008; Coulthard, 2010; Blackwell, 2012). Svartvik's 1968 article analyses four statements

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

allegedly made by Timothy Evans to authorities in which he admits to murdering his wife and baby daughter in 1949. However, following Evans' death, it was discovered that he was innocent of the murder of his young daughter and that his neighbour John Christie was guilty of that crime (and others) (Coulthard, 2004; Turell, 2008). Christie was convicted of killing at least five persons and sentenced to death in 1953, three years after Evans' execution. Evans' comments to the police were re-examined in 1966, thirteen years after difficulties arose in 1953 as a result of his allegations. It was contended that two of the statements were untrue and that their language did not match that of an analphabet (Evans could neither read nor write). Evans' language usage is studied from a linguistic standpoint in Svartvik's work. Svartvik (1968) emphasised that he conducted his analysis only to evaluate whether Evans's assertions could be considered genuine, and that the results should not be considered a legal judgement. Svartvik contrasted particular word choices and phrases in the written statements with those used by Evans throughout the hearing, and found that the wording in the statements was far more formal than the language Evans used at his trial. However, Svartvik's analysis could only ascertain that a certain percentage of the assertions were fake; he was unable to identify with certainty whether the rest were fabricated or authentic.

An interesting putative use of forensic linguistics may be traced back to a letter written by the British logician Augustus de Morgan in 1851. In this letter to a friend (Hockey, s.a), De Morgan suggests that it may be feasible to identify the writers of the biblical books by focusing on word length as a distinguishing stylistic trait (Kotzé, 2007). According to Kotzé (2007), forensic linguistics presumably began with De Morgan's proposal. However, no additional developments of De Morgan's concept were documented for over thirty years, until Mendenhall picked up the theme in 1887. He used word and sentence length as indicators of authorship of the works of Bacon, Marlowe and Shakespeare, basing his analysis on De Morgan's suggestion.

The use of sentence length and word length as authorship indicators was later disproven as a reliable indicator of authorship by Smith (1983), who discovered that when one compares the works of different authors within the same literary genre, the distribution of word lengths in the texts is so similar that the same person appears to be the author of all the texts (Holmes, 1994). Nonetheless, Mendenhall's research resulted in the creation of further techniques for author identification (Holmes, 1994; Holmes, 1998; Olsson, 2004).

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

Mosteller and Wallace's (1964) examination of the Federalist Papers is one of the most well-known forensic-linguistic analyses and is considered by some scholars to be the first serious forensic-linguistic analysis of texts that sought to discover the author of the works (Broeders, 2001; Mikros, 2012; Stamatatos, 2009). The Federalist Papers consist of 85 documents published anonymously between 1787 and 1788. The objective of the documents was to convince New Yorkers to accept the new United States Constitution (Holmes, 1998; Coulthard & Johnson, 2007; Kotzé, 2007). Three authors, Alexander Hamilton, James Madison and John Jay, claimed authorship of the documents. Mosteller and Wallace (1964) studied the texts to discover supposed idiolectic traits in an effort to resolve the issue. First, a collection of texts from each author was reviewed to discover if they used idiolectic terms or phrase constructs. In addition, they focused on the frequency of function terms in each text. Koppel et al. (2009, p. 5) refer to this analysis as a 'multivariate analysis approach' that 'heralded a new set of methods for stylometric authorship attribution, based on combining information from multiple textual clues'. According to Holmes (1994), the analysis exhibited striking similarities to Mendenhall's (1887) research. Following the identification of thirty idiolectic traits of each author's texts, Mosteller and Wallace (1964) determined that Madison was the author of the 12 problematic texts by comparing these traits to the linguistic features of the 12 documents.

This discovery is consistent with findings by historians who also attribute the 12 documents to James Madison. Because the author identification of the Federalist Papers was so effective, it is customary practice to continue testing new author identification methods and hypotheses on copies of the Federalist Papers (Holmes, 1998; Coulthard & Johnson, 2007; Stamatatos, 2009). However, Stamatatos (2009) emphasises that the Federalist Papers represent an ideal and an uncommon forensic-linguistic scenario, in that it had few potential writers and extensive texts with which to work. Linguists who test their procedures on these papers should bear this in mind, as forensic-linguistic conditions in the real world are likely to be different.

Grieve (2005) and Schulstad et al. (2012) offer an additional origin date of author identification as an area of enquiry. They believe that author identification methods were employed to determine the authors of literary texts as early as the 1700s. The investigations done by Grieve (2005) and Schulstad et al. (2012) are also regarded as pioneering research in quantitative analysis (termed stylometry in forensic linguistics). According to Edmond Malone's 1787 investigation, Shakespeare was not the author of any of the three passages comprising the text of the play *Henry VI*. Grieve (2005) states that Malone based this conclusion on the fact that

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

the author of Henry VI typically ends each sentence at the end of a line of poetry, rather than using enjambment – the overflow of a sentence from one line of verse to the next. In addition, Malone stated that the author rarely rhymed the final syllables of poem lines. These are atypical metrical aspects of Shakespeare's plays.

In the United Kingdom, the earliest investigations of forensic linguistics date back to 1985. This indicates that forensic linguistics gained popularity in Britain considerably later than in the United States. The majority of these ground-breaking investigations were conducted in Birmingham, where linguists testified in court trials involving handwriting analysis and authorship identification. Their testimony involved written and verbal texts (Turell, 2008).

Nonetheless, the consensus is that the subject of forensic linguistics became more formally established and recognised as a field of enquiry and study in the early 1990s. Between 1988 and 1992, there was a significant increase in interest in forensic linguistics, resulting in the organisation of a number of seminars in Britain and Germany. These lectures were attended by delegates from Australia, Brazil, Holland, Greece and Ukraine. As the subject evolved, conferences were eventually organised in Australia (1995) and the United States (in 1997). In 1993, the International Association of Forensic Linguistics (IAFL) was established, while the International Association of Forensic Phonetics (IAFP) and the International Journal of Speech, Language, and Law were established in 1994. These conferences and the formation of the aforementioned groups demonstrate that forensic linguistics has become an international topic of study and that individuals from many nations undertake research in the field and use their findings as courtroom evidence. In the years that followed, the number of forensic linguistics papers increased dramatically, as did the range of research topics. Papers were produced by Solan (1993), Levi (1994), McMenamin (1994, 2002), Stygall (1995), Kurzon (1997), Hanlein (1999), Elrich (2001), Foster (2002), Alcaraz and Hughes (2002), Rose (2002), Cotterill (2003), Gibbons (2003), Heffer (2005), Coulthard and Johnson (2007), Kniffka (2007), Eades (1995, 2008), Shuy (1993), Johnson and Coulthard (2010) and Turell (2008).

Members of the forensic linguistics community have been able to voice their thoughts on a variety of issues and topics through the creation of websites and online discussion forums as the discipline has grown in prominence. In addition, so-called online 'forensic-linguistic laboratories' and language and law websites have been formed. By the end of the 20th century, the area was well established and undergraduate and/or graduate courses in forensic linguistics

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

were offered at universities in the United States, Britain, Australia, China, Finland, Germany and Japan, among others (Blackwell, 2012; McMenamin 2002).

Forensic linguistics is a growing field in South Africa, and it is now being taught at several universities, such as the University of the Western Cape and North West University. These universities offer year modules in honours degrees, where in the past, the subject was studied only as part of applied linguistics. Akademia, a college in Johannesburg, also now offer 'An Introduction to Forensic Linguistics' as a diploma. The course is ideal for law students from NQF 6 level upwards, legal practitioners, criminologists, police officers, journalists, language practitioners, linguists, psychologists and behaviour analysts. This introductory course is taught by leading researchers in linguistics in South Africa: Dr Annelise de Vries, Dr Karien van den Berg, Dr Zakeera Docrat and Karien Brits. However, the expertise of forensic linguists is still not commonly requested by the courts, since the judicial system and police service do not completely understand what the field can offer. In the past six years there has been a rise in the number of conferences, colloquiums and round tables on the subject, in which researchers from across Southern Africa present and share their knowledge, creating space for young researchers to learn about the field in other countries.

Below, the few researchers and academics who have published research in forensic linguistics in South Africa are briefly discussed.

Moeketsi has contributed to the field of forensic linguistics in South Africa and in 1997 and 1999 published the following article and book, respectively: 'Of African languages and forensic linguistics: The South African multicultural criminal courtroom' and *Discourse in a multilingual and multicultural courtroom: A court interpreter's guide*. The focus of these texts was mainly language use in South African courts. Moeketsi aimed to determine whether communication between suspects and legal entities was effective in a multicultural context, as well as how the use of language in courts could intimidate suspects and affect proceedings.

Taylor (1998) published an article entitled 'Addressing the insane language of the law', which focuses on language use and its legal implications. In 2006, the same topic was investigated by Reddy and Potgieter in their article 'Real men stand up for the truth: Discursive meanings in the Jacob Zuma rape trial'. Lombard and Carney (2011) conducted a similar enquiry with their article 'Die wenslikheid van Afrikaans as vaktaal vir regstudente', as did Carney in 2012 in

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

‘n Forensies-semantiese beskouing van die woordgebruik “onkoste” in die hof saak *Commissioner for South African Revenue Service vs. Labat Africa Limited*’.

In 2002, Thetela’s article ‘Sex discourses and gender constructions in Southern Sotho: A case study of police interviews of rape/sexual assault victims’ was published in *Southern African Linguistics and Applied Language Studies*, comprising a case study of police interviews of rape and sexual assault victims. Thetela focuses specifically on gender studies in the legal system, but views this phenomenon from an African (specifically Southern Sotho) perspective. Thetela describes the research itself as a focus on sex discourse or ‘talk about sex’ and looks at how gender relations and identity are shaped in social systems through such discourse.

Sanderson (2007) examined the field of trademark disputes, which have also received little attention to date. The article is titled ‘Linguistic analysis of competing trademarks’.

In his article, ‘The case for cyber forensic linguistics’, Klopper (2009) investigates forensic linguistics from the point of view of computer science, which is also a little known field of research in South Africa.

Other researchers who are actively working to grow the field and make South Africa known in the body of knowledge are: Professor Monwabisi K. Ralarala, Professor Russell Kaschula, Doctor Zakeera Docrat and Doctor Annelise de Vries, with publications such as:

- ‘A compromise of rights, rights of language and rights to a language in Eugene Terre’Blanche’s (ET) trial within a trial: Evidence lost in translation’ (2012) (Monwabisi K Ralarala);
- ‘“Meaning rests in people not in words”: Linguistic and cultural challenges in a diverse South African legal system’ (2013) (Monwabisi K Ralarala);
- ‘Transpreters’ translations of complainants’ narratives as evidence: Whose version goes to court?’ (2014) (Monwabisi K Ralarala);
- ‘An analysis of critical “voices” and “styles” in transpreters’ translations of complainants’ narratives’ (2016) (Monwabisi K Ralarala);
- *New frontiers in forensic linguistics: Themes and perspectives in language and law in Africa and beyond* (2019) (Monwabisi K Ralarala, Russell Kaschula and Georgina Heydon);

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

- *Language and the law: Global perspectives in forensic linguistics from Africa and beyond* (2022) (Monwabisi K Ralarala, Russell Kaschula and Georgina Heydon);
- *A handbook on legal languages and the quest for linguistic equality in South Africa and beyond* (2021) (Zakeera Docrat, Russell Kaschula and Monwabisi K Ralarala);
- ‘The role of African languages in the South African legal system: Towards a transformative agenda’ (2018) (Zakeera Docrat);
- ‘A review of linguistic qualifications and training for legal professionals and judicial officers: A call for linguistic equality in South Africa’s legal profession’ (2022) (Zakeera Docrat); and
- ‘Multilingualism in the South African legal system’ (2019) (Zakeera Docrat and Annelise de Vries).

From the above, it is clear that research in forensic linguistics has been advancing in South Africa, with most focusing on the legal implications of language. Furthermore, the majority of studies are in English. For an overview of research done in the sub-category of authorship analysis, see Section 2.3.

2.2.2 Applied linguistic vs theoretical linguistics

As noted previously, forensic linguistics falls within the field of applied linguistics. However, theoretical linguistics still plays a role in linguistic forensic investigations. McMenamin (2002, p. 62) describes forensic linguistics as ‘one of several developing areas in applied linguistics that employs the scientific study of language to answer forensic difficulties’. For this reason, it is vital to describe the differences between applied linguistics and theoretical linguistics.

Applied linguistics derives from conceptions of second language teaching. As a result of advancements in the teaching of language skills, applied linguistics has grown to encompass additional fields of study. Language teaching has come to be regarded as far more than the teaching of grammar conventions, prompting theorists to begin incorporating sociology and the social sciences into language teaching. Wei (2011, p. 7) defines modern applied linguistics as follows:

[...] a broad field of study of language learning and language use by different learner and user groups as well as wider social issues such as language planning, language ideology and language and social (dis)advantage. It is no longer focused on applying any specific linguistic theory or model, but on developing a critical perspective on language in everyday social life.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

Methodologically, applied linguistics has adopted a broad-based discourse analysis, complemented by multimodality analysis.

In addition to second language instruction, applied linguistics draws from language planning, a consideration of intercultural stereotypes, forensic linguistics, media discourse and deaf studies, among many other disciplines of study.

Wei (2011) cautions against losing sight of the language structure itself and contends that applied linguistics is separate from sociology, economics, politics, and law, since it focuses on language. Wei states, 'At its foundation, applied linguistics requires a coherent theory of language, whether derived from linguistics or another discipline, a set of rigorous descriptive techniques for dealing with language, and a body of research pertinent to language practice' (Wei, 2011, p. 2).

This emphasis on language as the foundation of forensic linguistics requires that forensic linguists have an understanding of theoretical linguistics. As implied by the term, applied linguistics is the application of theoretical linguistics' expertise in forensic-linguistic investigations. Olsson (2008) notes that the application of theory in forensic linguistics is not synonymous with application in, for example, applied statistics. In the latter, a theory 'underpinning a specific discipline to the practice of that science' is applied, but in forensic linguistics, linguistic knowledge is used within a specific context – the legal environment.

McMenamin (2002, p. 57) provides a helpful definition of what could be deemed theoretical linguistics:

Linguistics is about understanding the system of language. The aims of linguistic science are theoretical insofar as linguists discover the underlying rules and patterns of language and then describe them in the languages of the world. Linguists look for language characteristics that are present in all languages (universals), as well as features found only in certain language families or individual languages.

Theoretical linguistics is concerned with the language microsystem, and is the study of the language system, which comprises features such as the language's lexicon, syntax, morphology, phonology, phonetics and pragmatics (Johnson & Johnson, 1999). In theoretical linguistics, the linguist attempts to determine, among other things, what patterns and language norms exist within a given language. This understanding of a language system enables the linguist to recognise, within the sub-category of author identification, for instance, distinct deviations and variances between the writing styles of different writers. However, a knowledge

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

of theoretical linguistics is also used in other subfields of forensic linguistics, such as forensic phonetics.

2.3 Authorship analysis in South Africa: An emerging sub-category

In South Africa, research in the field of forensic linguistics, particularly authorship analysis, is extremely uncommon, despite indications that forensic linguistics is a developing area of investigation. It has already been stated that only a small number of people in South Africa have conducted research or are currently working in this field. The few scholars and researchers who have published research on authorship analysis are discussed briefly below.

Hubbard and Kotzé are both professors in this discipline. Hubbard has produced a number of articles on forensic linguistics. These include ‘Errors in court: A forensic application of error analysis’ (1994), ‘Linguistic fingerprinting?’ and ‘A case study in forensic stylometrics’ (1995). In addition, Hubbard presented an unpublished study on stylometry: ‘Stylometric and error analysis in the context of a style shift in abusive e-mail texts’ (2009). Only a small number of Hubbard’s essays and publications are solely concerned with author identification, with the analyses based only on English texts.

‘“Die vangnet van die word”: Forensic-linguistic evidence in a libel case’ (2007) and ‘Author identification from competing views in forensic linguistics’ (2010) are the titles of Kotzé’s articles. The subject of both of these articles is author identification in English writings.

Van den Berg has published a few works on hate speech, including the 2019 book chapter, ‘A case of crying wolf? A linguistic approach to evaluating hate speech allegations as linguistic acts of violence’. She also published a book chapter in 2019 with Surmon, adapted from Ms Surmon’s Master’s thesis, ‘The act of threatening: Applying speech act theory to threat texts’.

Examples of research conducted in Afrikaans in the sub-discipline of authorship analysis are a doctoral thesis by Anneen Church-Fleischmann in 2020, ‘’n Korpusgebaseerde ondersoek na kohesiepatrone as moontlike stilistiese kenmerk van outeurstyl’; and a master’s thesis by Lezandra Grundlingh in 2015 titled ‘Outeuridentifikasie: ’n Forensies-taalkundige ondersoek na Afrikaanse SMS-taal’. Surmon’s (2013) masters dissertation targets a similar topic to the current research and examines authorship identification on the social network, Facebook. The title of the dissertation is ‘Investigating the use of forensic stylometric analysis to determine authorship on a publicly accessible social networking site (Facebook)’.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

Kotzé and De Vries are adding to the field through their work on an authorship analysis case where disputed texts were involved. They were approached by a client who suspected an author of sending defamatory texts. Kotzé conducted a theoretical linguistic analysis, while De Vries conducted a socio-linguistic analysis to determine a linguistic profile. The matters of the case cannot be discussed in detail, but it is noteworthy that based on both these analyses, the texts in question were attributed to an author.

Clearly, authorship analysis research has been conducted in South Africa, but the majority of studies have focused on the legal aspects of language use. In addition, much of the research was not taken further; only three of the above researchers are still engaged in research on authorship analysis.

2.4 Authorship attribution/identification

2.4.1 Where it started

The roots of author identification are closely tied to the origins of forensic linguistics as a field. The initial forensic investigations mentioned in Section 2.1.1 may be summarised as follows:

- Edmond Malone's (1787) investigation that sought to demonstrate that William Shakespeare was not the author of Henry VI.
- Augustus de Morgan's 1851 letter proposing that the writers of the individual Bible books could be identified based on sentence and word length.
- Mendenhall's (1887) attempt to affirm or deny Bacon, Marlowe and Shakespeare as the writers of many works, using De Morgan's concept.
- Mosteller and Wallace's (1964) attribution of the Federalist Papers' writer.
- Jan Svartvik's (1968) investigation in which he attempted to discover whether Timothy Evans' written confession of having killed his wife and children was based on fact.

The broad discipline of forensic linguistics had not yet been defined as a field of study when these initial author identification investigations were conducted. It was only when forensic linguistics developed to incorporate additional disciplines of study, such as forensic phonetics, language rights and language use in the legal environment, that sub-categories emerged such

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

as author identification. When these areas of interest developed, the umbrella term for the discipline became known as ‘forensic linguistics’.

The methodological advancements in author identification have been remarkable. The approaches such as those employed by Mendenhall (1887) were later shown to be flawed (Smith, 1983); as a result, scholars have attempted to develop newer and more precise statistical methods. However, many of the early techniques were not regarded as legitimate (Juola, 2006). Several of these techniques are outlined by Holmes (1994) and Stamatatos (2009), and a few are described in greater detail below. One needs to keep in mind that variations in the approach used for author identification to date may be ascribed to forensic linguists’ inability to find the appropriate corpus size or text length for effective author identification in each case. In research, however, there are ideas regarding what constitutes the ideal corpus. Stamatatos (2009) suggests that an optimal analysis is only possible when the genre and subject are under absolute control. However, this may be considered an unrealistic expectation. The relevant facts and texts in actual cases simply do not permit such control (Juola, 2006).

2.4.2 Research in the field of author identification

As is evident in the preceding discussion, forensic linguistics forms the broad area of several research studies conducted in recent years. In the more specific subdiscipline of author identification, researchers have also conducted a range of investigations.

There is growing interest in the notion that each person has a distinct writing and speaking style, and that the distinctions may constitute individual idiolects. Since the late 1990s, the incorporation of the term ‘idiolect’ in author identification research has expanded dramatically. Numerous forensic linguists have deemed idiolect to be an essential concept in the area of forensic linguistics, and as a result, idiolect has been the subject of numerous investigations. As a topic in author identification, it appears to be essential to the work of a number of prominent forensic linguists.

2.4.2.1 A focus on idiolect

In one of his works titled ‘Author identification, idiolect, and linguistic uniqueness’, Coulthard (2004) analyses the prospect of using idiolect to identify authors and detect plagiarism. Coulthard maintains that individual idiolect can result in a distinguishable difference between the works of an original author and the work of someone who has used part of that work in

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

their own text, claiming it all as their own. In addition to defining idiolect, Coulthard (2004) notes that the proportion of individual words (or lexical categories and signals) that are the same in two or more texts should be emphasised. This proportion is a reliable measure of which text is the original and which contains plagiarised content.

McMenamin (2002) explores author identification and the significance of idiolect in author identification. McMenamin (2002) discusses forensic stylistics, among other topics, in his book *Forensic Linguistics: Advances in forensic stylistics*. Here, written texts are analysed through an identification of language usage trends and variety in a particular document (2002). According to McMenamin (2002), such patterns and variations are part of the author's idiolect, and through identification of idiolect, authorship may be assigned to various texts. McMenamin (2002) mentions a scale created by the Scientific Working Group for Forensic Document Examination (SWGDOC). This scale consists of nine points, each of which represents a percentage of similarity between the suspect text and the known text. A score of nine on the scale indicates that all requirements are met, resulting in a positive identification. In other words, all the suspect text's properties match those of the well-known text, which indicates that the author of the well-known text and the suspect text are one and the same. Six, seven, and eight indicate that the author of the suspect text has been positively recognised, although there are several discrepancies between the suspect text and the known text. A score of five indicates that no conclusion may be formed; scores of two, three and four imply a significant likelihood that the author of the well-known text is not the author of the suspect text. A score of one indicates that the author of the known text cannot be the creator of the suspect text.

Similar to Coulthard, McMenamin (2002) notes that the use of forensic stylistics raises various difficulties that call into question the validity of a forensic-stylistic study. These problems include, among others, whether stylistics is a sufficiently established field to yield valid results in forensic linguistics and whether a standard for diversity in language use exists and is attainable. McMenamin (2002) argues that ways to make author identification more scientific are continually being developed and evaluated, and that any social, geographical or situational norm may be formed and used to explain variations in written language.

Olsson (2004; 2008) conducted numerous experiments in author identification, and approaches idiolect from a unique perspective. The term 'linguistic fingerprint' (which allows forensic linguists to identify the authors of works) is occasionally used as a synonym for 'idiolect'

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

(Coulthard, 2004; Brennan et al., 2012), and Olsson (2004) argues that this is problematic. Although Olsson mentions the use of this word by a number of scholars, he argues that the concept of a linguistic fingerprint is a fiction, adding, ‘the lack of evidence for its existence is striking’ (Olsson, 2004, p. 31). The argument against the use of the phrase ‘linguistic fingerprint’ is elaborated upon in Section 2.7.3, in which it is shown that several scholars endorse Olsson’s position.

Chaski (2005) holds that idiolect exists and can be determined by stylometric analysis. Chaski’s stylometric analyses are based on ‘easily computable and countable linguistic variables, such as word length, phrase length, sentence length, vocabulary frequency, and word length distribution’. However, Chaski primarily focuses on author identification in digital testimony writings. These are computer-generated writings linked to crimes such as murder, financial crimes, threats and identity theft. In some instances, a person commits crimes using digital texts while posing as someone else. Author identification in text messages may be considered digital evidence to some degree, given that criminals might use text communications to conceal their identity after committing crimes. Based on her examination of three cases, Chaski concludes that author identification cannot function as a separate field of enquiry, but must be used in conjunction with well-established forensic procedures, such as biometric analysis of the keyboard user, to determine the author with certainty. The final objective of Chaski’s (2005) enquiry is to evaluate a quantitative method that she believes identifies authors with 95% accuracy. This technique is known as syntactic analysis. According to Chaski (2005, p. 3), syntactic analysis differs from other stylometric analyses in that it is ‘linguistically sophisticated and grounded in linguistic theory’.

2.4.2.2 A focus on short texts

Author identification in longer digital texts is possible and fairly frequently used, while author identification in brief digital texts such as text messages and other emerging forms of communication such as tweets and status updates is less common. Several researchers, such as Chaski, concentrate on digital texts, particularly shorter digital writings. Grant (2010; 2012) is a researcher whose work on author identification in brief texts explores a variety of possibilities and issues. An article by Grant titled ‘Txt 4n6: Idiolect-free authorship analysis?’ (2010) addresses the question of authorship identification reliability, specifically with respect to idiolect. Grant (2010) focuses on two difficult features of author identification: the length of texts (especially brief digital texts) and the existence of idiolect. This author contends that

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

idiolect is not as easily observable as some academics believe, and that the reason for the use of idiolect in any given text analysis must be explicable. According to Grant (2010), the assertion of idiolect's value is particularly challenging for linguists. To tackle this issue, Grant (2010) advises that the linguist employ a mixed approach in which theoretical linguistics and cognitive linguistics explain idiolect in a specific text. The subject of cognitive linguistics is discussed further in Section 2.7.4.1

McLeod and Grant (2012) examine the issues associated with succinct texts and the popular notion that author identification methods should be modified for shorter texts. The majority of author identification techniques employed in prior studies are based on longer, mostly literary works. McLeod and Grant (2012) hold that similar techniques are not appropriate for evaluating shorter texts. These authors (employed an adaptation of Jaccard's co-efficient approach after analysing and comparing the methods of other researchers. The method is now known as the Delta-S method. According to McLeod and Grant (2012, p. 221), 'The method described here enhances the state of the art in terms of the message size for which authorship analysis may be performed.' However, they agree that further enhancements are possible and necessary in order to increase the method's precision.

It is evident from the above discussion that authorship studies encompass a wide variety of texts, both digital and printed. Obviously, the discussion does not include all the studies that have been conducted to date on author identification, but the selection does show that two basic topics of importance in current research have already emerged – idiolect and compact (digital) texts. Both have already been investigated on multiple occasions. It is evident from the literature that the term 'idiolect' presents a number of difficulties, as it is not possible to establish that an idiolect is present in every case. The number and length of texts have an important bearing on the precision of results regarding the presence of idiolect. Small quantities of short texts can mislead the researcher into believing that an idiolect is present in each dataset, but the idiolect that the researcher detects may in fact be a consequence of chance and may evaporate once the researcher has access to more texts. However, success has been gained in author identification studies of brief texts, particularly when n-gram and log-likelihood approaches are used. It is essential to keep in mind that each circumstance is unique and that certain strategies will be more effective in some scenarios than in others. In order to maximise the accuracy of the results, it is recommended that author identification studies include multiple methods, particularly in the study of brief texts (McLeod & Grant, 2012).

2.5 Author identification and the court

Several courts (including those in the United States, parts of Australia, and England) do not consider evidence gathered by forensic linguistic methods to be admissible unless it has been corroborated. This means that the evidence gathered by the forensic linguist may be released, but there is no guarantee that it will be used or acknowledged in court. Even if the forensic linguist testifies as an expert witness, their testimony may be dismissed. Chaski (2001) notes that forensic linguistic evidence is only accepted if the procedures used to identify each element in the text, such as the author, have been scientifically evaluated. This means that the methods of identifying the author must have been peer reviewed and published, making the method scientifically accepted. This is simply one of the conditions for court-admissible evidence.

2.5.1 The forensic linguist's role in court cases

According to Coulthard (2010), there are two categories of witness in a court case. The first kind of witness is personally connected to the case or individuals in it, and the second kind has no personal connection to the case. This witness is a specialist witness. When linguistic forensic experts testify, they do so as expert witnesses.

The court or an investigating team may contact linguists for a variety of reasons. Coulthard (2010) discusses instances in which linguists have served as expert witnesses. Examples include cases about trademark issues (Shuy, 2002; Gibbons, 2003) and cases demonstrating the legitimacy of texts in two murder cases (Turell, 2004; Coulthard, 2002). Olsson (2008, p. 4) describes the services provided by linguists to the court as 'linguistic intelligence work'. Such work includes an examination of text messages, threats and ransom notes. Linguists may also be tasked with analysing alleged suicide letters in order to evaluate whether or not the content is genuine. The police may also request the linguist's opinion on a specific document or audiotape. In court, it is unlikely that the linguist's opinions regarding the veracity of a text will be regarded as reliable evidence. Consequently, linguistic studies are typically restricted to the initial (investigative) phase of an investigation (Olsson, 2008).

During the trial phase of an enquiry, linguists are typically called upon to perform analysis pertaining to author identification, meaning interpretation, written threats, or the provenance and composition of writings. The criminal or civil character of the case will decide the admissibility of such evidence (Olsson, 2008). Typically, if a suspect or accused is found

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

guilty, the defence will file an appeal. During this phase, the linguist may be relied upon to resolve a dispute over the phrasing, interpretation or authorship of a confession or statement. A linguist may also need to verify a new interpretation of a forensic text (such as a suicide note or ransom note) that comes to light after the conviction (Olsson, 2008).

Solan (2010, p. 400) cautions the forensic linguist to be aware of two characteristics of court situations with which they may be faced:

[...] the brutality of the adversarial system, including snide and personal attacks on individuals working within standard scientific paradigms, and a broad concern that the forensic identification sciences lack adequate scientific foundation.

In an effort to address the matter of personal and discipline-related attacks, Solan (2010) offers proficiency assessments for forensic linguists. Since courts are where forensic linguists spend much of their time, the methodology employed in the creation of such competency exams should be evaluated outside of court. Proficiency exams (further discussed in Section 2.5.2.1) should be complemented by the establishment of protocols and the preparation of reports showing the linguistic knowledge and reliability of the forensic linguist. According to Solan (2010), the design of proficiency examinations is challenging since it is difficult to create content related to actual, everyday forensic issues. When there are only two potential authors of a document, it is relatively easy to determine which is the true author, since the rejection of one author will imply that the other is the author. In other situations, the forensic linguist must assess whether a particular suspect is the author of a certain text; in such cases, there are no other suspects, and thus the question arises of how many potential authors must be eliminated before the linguist can definitively state that the suspect produced the text. For proficiency examinations to be meaningful, these issues must be resolved.

2.5.2 Forensic-linguistic evidence in courts: A brief overview

2.5.2.1 The United States of America

According to the courts of the United States, linguistic forensic evidence is frequently disregarded because it fails the Daubert standard. According to the Daubert test, in order for evidence to be considered valid, it must satisfy the following conditions: (i) The expert witness must have adequate specialised expertise (in this case, forensic linguistics is referred). This expertise should consist of field experience, training and instruction. In addition, the witness must be a respected member of his or her academic or other peer group. (ii) The evidence-

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

gathering method must be empirically evaluated. In addition, it must be refutable (Chaski, 2001). (iii) The technique must have already been peer reviewed and published. iv) The expert must be able to specify the technique's margin of error. (iv) The technique must be able to be communicated to the court with sufficient clarity so that the main notion behind the technique may be comprehended by everybody (those in court and the general public) (Olsson, 2004). In the United States, not all courts recognise the Daubert test or examine all of its requirements. However, when the Daubert test is recognised, most of its requirements must be met for evidence to be considered legitimate. Chaski (2001) cites *United States v. Van Wyk*, 83 F. Supp. 2d 515, D.N.J., 2000, in which the court determined that the State's evidence met only one of the Daubert test requirements. For this reason, the court decided not to accept the State's testimony (Chaski, 2001). Although Fitzgerald (an expert witness in the case) employed a methodology that may be subject to testing, neither Fitzgerald nor the US Government had been able to identify a known rate of error, establish the number of samples required for an expert to reach a conclusion regarding the probability of authorship, or identify a peer review that is meaningful. In addition, as argued by the defendant in the above case, there is no globally accepted certification requirement for forensic stylistics experts.

The Daubert test replaces the Frye test, which was previously used by courts to evaluate the reliability of evidence. The Frye test, also known as the general acceptance test, assumed that 'if a method was accepted by the scientific community, it might be considered admissible in court' (Olsson, 2004, p. 41).

In the United States, there are now three kinds of author identification techniques whose results are acceptable in court if they have been empirically tested. These strategies are stated as follows by Chaski (2001, p. 2–3):

In the first group there are two techniques – syntactically classified punctuation and syntactic analysis of phrase structure – which withstand the scrutiny of experimental testing and statistical analysis. [...] In the second group are several techniques – sentential complexity, vocabulary richness, readability, content analysis – which quantify linguistic patterns, and are amenable to statistical testing. [...] In the third group are 'forensic stylistic' techniques – spelling errors, punctuation errors, word form errors, grammatical errors – which are rooted in handwriting identification and prescriptive grammar.

According to Mitchell (2008), the Council for the Registration of Forensic Practitioners recognised forensic linguistics as a specialised field on 1 September 2008. The recognition strengthened confidence in forensic linguistics (Mitchell, 2008). Mitchell (2008) cites Grant, a

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

professor of forensic linguistics at Aston University, who argues that this recognition demonstrates that forensic linguistics is increasingly considered a scientific area.

The National Research Council of the National Academies (NRC) published a report on the status of forensic identification in 2009. The content of this paper, titled 'Strengthening forensic science in the United States: A path forward', is relevant to all forensic techniques. Solan (2010, p. 398) mentions the following observations from the report (NRC 2009, p. S-7) that are also relevant to forensic-linguistic evidence:

Two very important questions should underlie the law's admission of and reliance upon forensic evidence in criminal trials: (1) the extent to which a particular forensic discipline is founded on a reliable scientific methodology that gives it the capacity to accurately analyse evidence and report findings and (2) the extent to which practitioners in a particular forensic discipline rely on human interpretation that could be tainted by error, the threat of bias, or the absence of sound operational procedures and robust performance standards. ... Unfortunately, these important questions do not always produce satisfactory answers in judicial decisions pertaining to the admissibility of forensic scientific evidence proffered in criminal trials.

The study advises that a substantial amount of research is required in a specific study area in order to determine the limitations of the methods or procedures and the potential for variation and bias. In this regard Grant (2008) proposed the creation a dedicated language database to quantify linguistic data, referred to in Section 2.5.2.2. Solan (2010) cites Olsson (2004), who argues that further research should be conducted in the field of author identification of text messages in order to increase the credibility of author identification results in this particular field of study.

Solan (2010) argues that variations in writing styles among individuals, and even within an individual's own writing style, are of major importance in author identification concerns. This variety should be taken into account when one is establishing the predictive power of these conventions for co-authorship and non-authorship, based on existing corpora or corpora obtained for research purposes.

In addition, it is proposed that the sub-categories of forensic linguistics be based on the sorts of documents analysed during investigations, and that the arguments for and against particular positions be standardised so that the legal community may reach consensus (Solan, 2010).

It would also be extremely beneficial for the future of forensic linguistics if forensic linguists who are regularly involved in a certain sort of forensic enquiry documented their experiences

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

and methods. These could then be used in the future by other forensic linguists. According to a report published by the National Research Council (NRC, 2009), research of this kind is currently insufficient, particularly in forensic fields that rely on subjective judgements of matching traits. In addition, the report recommends that the forensic professions create stringent rules to prevent subjective interpretations on texts. Solan (2010) also argues that programmes for research and assessment should be devised that adhere to stringent regulations and criteria.

Solan (2010) cautions those working in forensic domains and doing forensic analysis against confirmation bias. A typical example of this bias occurs when the police notify the forensic linguist that they are certain the suspect is guilty, and that the linguist merely needs to confirm this. In such a situation, it is crucial that linguists maintain objectivity in their studies and findings, as the conclusion reached will be investigated by others.

2.5.2.2 Australia

In Australian courts, linguistic forensic evidence is not required to meet Daubert standards, but judges emphasise that evidence must be relevant and credible. This guideline is known as the reliability and relevance rule (Olsson, 2004). Expert witnesses must guarantee that they can qualify the conclusions, noting where they think the evidence to be insufficient or erroneous, and they must also justify their opinions. This indicates that Australia's courts do consider the opinions of expert witnesses. Olsson (2004) notes, however, that there are still several Australian courts that are sceptical of linguistic evidence owing to the widespread belief that each individual has the linguistic knowledge necessary to form their own opinions on a text and that the study of language is not a technical field of study. This is a worldwide issue, and in this regard, Grant (2008) states that he has created a dedicated language database to quantify linguistic data. The database contains over 8,000 texts, with each text statistically analysed. Grant (2008) recommends the use a database such as this because, as he states, it is essential for the forensic linguist to exhibit proficiency in the discipline. Moreover, this expertise must manifestly exceed that of the average juror.

2.5.2.3 England and Wales

The English and Welsh legal system is different from those of the United States and Australia. For some decades, there were concerns that English and Welsh law was becoming inaccessible to the general public and that the court system was in need of reform. In the early 1990s, Lord

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

Woolf was appointed to undertake reform of the civil court system. According to Woolf, the involvement of specialists in a court case incurs additional expenditure, which discourages the public from seeking clarification of matters through legal proceedings (Olsson, 2004). Woolf proposed that just one expert be nominated for both parties, and that in the event that a single expert cannot be chosen, the court should select the expert. According to Woolf, the rationale for this is that experts must be cost effective and that any conflicts between two or more experts should be resolved as soon as feasible. However, Woolf did not stipulate any scientific evidence requirements comparable to the Daubert standard (Olsson, 2004).

2.5.2.4 South Africa

In South Africa, as in the United States and Australia, there are rules regarding the reliability of electronic evidence. These rules are essential for the current research, which was conducted within this legal framework. In South African evidentiary law, the admissibility of electronic evidence is evaluated in three processes, according to Watney (2009).

First, it must be determined what form of electronic evidence is employed. The evidence may fall under the category of documentary evidence or real evidence. Documentary evidence implies that the evidentiary weight of a piece of evidence is decided by the credibility of the source. In other words, the data was entered by a person and a computer/machine did not sample or alter the data. The evidentiary weight of real evidence is decided by the machine's hardware and software reliability. This applies when a person submits data into a computer/machine and the computer/machine converts the data to a different format than that originally entered (Thiart, 2015). Real evidence is also deemed to include documents generated by a computer/machine without the participation of a person – where the user merely activates the electronic system, which then automatically records and stores data. The system's software therefore generates the data, rather than a person (Thiart, 2015). These two categories of electronic evidence – documentary and real – apply exclusively to evidence that can be categorised as papers, i.e., books, maps, plans, drawings, photographs, pamphlets, lists, letters or records (Watney, 2009; Thiart, 2015). Watney (2009, p. 8) states that the term 'document' can also apply to 'any device by which information is collected or retained'.

Second, it is essential to evaluate the initial authenticity or legitimacy of electronic evidence. Watney (2009, p. 8) provides the following description of this requirement:

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

The South African law of evidence requires that anyone who wants to use a document as evidence has to satisfy the court that it is authentic, in other words, that the document is what it claims to be. Due to its high degree of volatility, electronic evidence can easily be manipulated, altered or damaged after its creation and therefore authenticity must be proved.

Third, the evidentiary weight of the evidence must be established. The evidentiary weight of electronic evidence is determined by the following criteria: (i) the reliability of the way in which the data was generated, stored and communicated; (ii) the reliability of the way in which the data's integrity/completeness has been maintained; (iii) the way in which the data's creator has been identified; and (iv) any other relevant factors (Watney, 2009). Watney (2009) cites Hofman (2006), who believes that, in addition to the aforementioned criteria, the court should also use experts to explain the technical procedures underlying the acquisition, processing and storage of electronic data in order to facilitate decisions regarding the evidentiary weight of the evidence.

However, the validity of the evidence is considered only once it has been determined that the evidence was collected legally. There are instances in which evidence has been obtained in an unconstitutional manner; in some cases, such evidence has been admitted because, according to Watney (2009, p. 3), fairness requires, in some instances, that 'evidence, although unconstitutionally obtained, be admitted nonetheless'.

Regrettably, despite this stipulation that evidence must be collected legally, few mechanisms exist in South African law to oversee the collecting, preservation and presenting of electronic evidence in criminal trials (Watney, 2009). This indicates that South African law addressing electronic evidence is occasionally impeded.

Any forensic study of evidence must adhere to stringent restrictions, and it is precisely these constraints that may hinder the validity of author identification as a competent method in court proceedings. The fact that no author identification method is one hundred percent accurate is unquestionably the largest challenge (Holmes, 1994; Holmes, 1998; Chaski, 2005; Ishihara, 2011; McLeod & Grant, 2012). This is a problematic issue, as suspects are not supposed to be sentenced unless the evidence against them proves their guilt beyond a reasonable doubt. The second issue is that there are so many methods for determining the likely author of a text that it is impossible to establish which method produces the most accurate findings. For these reasons, author identification cannot be considered the most important evidence in a case; rather, it is deemed supplementary analysis or circumstantial evidence. However, despite the

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

lack of absolute confidence, author identification of texts has already led to convictions in a number of cases (Crystal, 2008; Grant, 2010; Blackwell, 2012).

As previously stated, it is essential to employ a dependable method in identification assessments of texts. Reliable procedures contribute to the reliability of evidence and, in some situations, strengthen the circumstantial evidence against a suspect. In this enquiry, the chosen method of analysis is stylometry. Although the accuracy of some stylometric methods is contested, stylometry is a popular analytical technique among forensic linguists, as it combines numerous different methodologies during a single investigation and has had high success rates in author identification investigations in a number of cases.

The following section examines stylometry as a method of authorship attribution in more detail, along with other methods of authorship analysis derived from three theoretical approaches.

2.6 Stylometry

Kotzé (2007, p. 388) defines stylometry as:

[...] a thorough quantitative analysis, by means of which the relative frequency of identical vocabulary items or word groups is compared. It is called a quantitative analysis because it is based on the quantification of textual features as a basis for further calculations, meaning that each and every word is recorded and to be counted. A number of calculations are then performed on the data, followed by statistical significance tests.

Stylometry consists mostly of two processes: the identification of characteristics and the statistical processing of these features using a classification method (Barry & Luna, 2012). These two components indicate that stylometric analysis is dual in nature. The linguist must first determine which textual characteristics will be selected for processing. Then, an algorithm must be devised to statistically process the traits, so that a determination may be made of how common or uncommon each of these is. These two strategies may, in fact, be viewed as two distinct forensic linguistics techniques. According to Kotzé (2007), the selection of textual aspects and their examination is referred to as stylistic analysis and is qualitative in nature; the measurement of these stylistic elements and the statistical tests performed on them is referred to as stylometry, and is quantitative in nature. It is crucial that the stylistic analysis and stylometric study of a text or texts complement one another. The stylistic analysis is a crucial first step, since it identifies the characteristics that will be examined by the stylometric study.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

In addition, the characteristics selected must meet certain parameters for the stylometric analysis to be beneficial. Grieve (2005) discusses the stylistic examination of texts for author identification and highlights the necessity of stylometric analysis as part of the technique.

A successful quantitative authorship attribution depends on the investigator's selection of a set of textual measurements whose values are relatively consistent across each possible author's writing sample and relatively variable across the set of possible authors [...].

Although stylometry was employed to identify the writers of literary texts as early as the eighteenth century, Schulstad et al. (2012, p. 1) assert that stylometry is no longer used exclusively for literary or historical objectives. These authors assert that stylometry in contemporary forensic linguistics has far broader application:

[...] it also has forensic applications. [...] More recent studies have used stylometry to determine the authorship of e-mails and online messages to counteract cybercrime. In addition to identifying an author, stylometry can also be used to detect multiple authors in a text (plagiarism) or to assign an author to a sociolinguistic category such as gender.

In his paper, Holmes (1998) discusses stylometric methods used in the past and describes a variety of stylometric procedures used to increase the precision of stylometric studies. The identification of function words (prepositions, conjunctions and article titles) is one of the most precise stylometric techniques used in the past. These words are usually employed by an author without much thought and are not context dependent. During a stylometric analysis, one determines how frequently certain function words occur in, say, every 1,000 words, and a corresponding frequency is calculated. The frequency of occurrence in one text may then be compared to the frequency of occurrence in other texts to assess the likelihood that the author of one text also authored others.

Ellegard (1962) used this technique to establish the authorship of the Junius Letters, and Mosteller and Wallace (1964) used it to determine the authors of the Federalist Papers. Brown-Jackson (2013) employed a similar strategy, among others, to show that J.K. Rowling is the author of the debut novel *The Cuckoo's Calling*. Rowling published under the pseudonym Robert Galbraith but was recognised as the author through the use of forensic-linguistic techniques such as the order of adjacent words, the sequence of characters, a calculation of the most common words in the text, and the author's predilection for long or short phrases (Zax, 2014).

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

According to Holmes (1998), principal component analysis and so-called neural network analysis produce the most accurate results in stylometry. However, enormous amounts of text are necessary for such analyses in order to generate accurate findings. In addition, the number of potential authors in an investigation might impair the accuracy of certain approaches. According to Luyckx and Daelemans (2011, p. 37), a high number of potential authors can even sway normally dependable procedures such as the detection of function terms:

[...] it may be correct to claim that distributions of function words are important markers of author identity, but the distribution of a particular function word, while useful to distinguish between one particular pair of authors, may be irrelevant when comparing another pair of authors.

In other words, the optimal condition for author identification is a small number of suspect authors and voluminous quantities of material. These conditions are rarely met in court case enquiries.

As previously stated, the premise of stylometry is that the core of each author's unique style may be captured by a certain number of quantitative criteria (Somers, 2008). These quantitative criteria are also called discriminators. Although a great number of style elements are chosen accidentally by the author, other elements are in fact chosen purposefully in the context and subject of the work. In other words, there are replicable features of each author's style.

Computer-based stylometry makes it easier to differentiate conscious style markers from unconscious style markers in the works of various authors, which is why this method is so popular (Somers, 2008). It is vital to remember that artificial intelligence currently dominates stylometry. Consequently, the human element of stylometry is becoming less relevant (Brennan et al., 2012). In the future, the systematic movement in stylometry toward a more computer-centric approach could bolster confidence in this technique and contribute to the recognition and admissibility of such studies as evidence in court. With the creation of ChatGPT⁵ (2022) and the 5th Industrial Revolution⁶ looming, the human component is being removed more and more from society and business, meaning that society has to adapt and

⁵ ChatGPT is an artificial-intelligence (AI) chatbot developed by OpenAI and launched in November 2022. It can write and debug computer programmes; compose music, teleplays, fairy tales and student essays; answer test questions (sometimes, depending on the test, at a level above the average human test-taker); write poetry and song lyrics; emulate a Linux system; simulate an entire chat room; play games like tic-tac-toe; and simulate an ATM (OpenAI, 2022).

⁶ The Fifth Industrial Revolution, or 5IR, encompasses the notion of harmonious human-machine collaborations, with a specific focus on the well-being of the multiple stakeholders (i.e., society, companies, employees and customers) (Regenesus Business School, 2020). It thus paves the way for a (r)evolution in thinking about and leveraging human-machine collaborations for greater societal well-being.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

create new technological advancements. This naturally affects the ways in which authorship identification is conducted through stylometric methods.

Although stylometry is frequently employed for author identification, scholars should be mindful of its limitations. As previously mentioned, no textual feature combination is a hundred percent accurate in the context of a stylometric study. According to a 1998 survey of stylometry, over a thousand features were already included in stylometric research in the 1990s (Schulstad et al., 2012; Zechner, 2013). These features consisted of lexical, syntactic, structural, idiosyncratic and context-specific characteristics. Lexical features refer to the number of words in the lexicon, whereas syntactic features refer to sentence patterns and the employment of particular function words. Structural features refer to the way the text is constructed and take into account elements such as the presence of paragraphs, the length of paragraphs, and the use of indentation. Idiosyncratic characteristics include intriguing aspects linked to the spelling of words or other features, and context-specific qualities pertaining to the use of context-specific terms (Schulstad et al., 2012). To guarantee that stylometry is as accurate as possible, a number of academics have concluded that the linguist should pay special attention to function words (grammatical words) and syntactic aspects, but should also examine semantic and lexical features such as discriminators (Holmes, 1998; Luyckx & Daelemans, 2011; Koppel & Schler, 2004). Schulstad et al. (2012) cite a study by Grieve (2005) in which it was determined that the repetition of function words, punctuation, bi-grammes (2-grammes), and tri-grammes (3-grammes) are the most effective authorship indicators in a stylometric examination. Grieve (2005) hypothesises that success in the use of this combination of factors is due to the fact that function words and punctuation marks signal how sentences are created, while the content of the text influences the use of n-grams.

Researchers should also keep in mind that successful stylometric analysis typically employs vast quantities of data. Luyckx and Daelemans (2011) regard the ideal data to be large texts for each author or multiple short texts for each author. In other words, the success of a stylometric analysis reduces as the amount of text available to the forensic linguist decreases.

Although the potential success of a stylometric study depends on the circumstances of the analysis and the quantity of text, the linguist can pose the following closed questions. These questions can aid the linguist in the author identification procedure.

- Did author A or author B pen/publish this text/document?

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

- If author A created this document, did he or she also create these documents?
- What is the likelihood that author A authored/produced this document?

The sub-categories of stylometry are stylochronometry and adversarial stylometry. The purpose of stylochronometry, according to Corney (2003), is to determine the chronological order of an author's works. Corney (2003) believes that whereas stylometry implies that each author has a distinctive style, stylochronometry assumes that this distinctive style will evolve over time. Corney (2003, p. 21) mentions researchers who have already conducted stylochronometry research and, based on their findings, makes the following observation:

The findings of these research suggest that an author's style can and does evolve with time. In these instances, the time span in question was greater than ten years. These results should be considered for any forensic investigations, and the known writings of any given author under enquiry should be sampled from a relatively brief period of time, such as one or two years.

The following section elaborates on the second sub-category, adversarial stylometry.

2.6.1 Adversarial stylometry

Adversarial or adversative stylometry may be viewed as a particular angle or perspective on stylometry. Adversarial stylometry challenges the premise that the author of a given text does not actively alter his or her writing style prior to producing the text (Brennan et al., 2012). The conventional idea is that the author of a book (or any other text) always uses his or her true writing style and never knowingly alters it. Adversarial stylometry challenges this assumption and holds that in fact authors may deliberately attempt to disguise their identity for various reasons. In adversarial stylometry, the linguist attempts to establish what approaches untrained individuals might employ to hide their writing style and the degree to which these techniques can trick existing stylometric tools. Brennan et al. (2012, p. 2) offers the following definition of adversarial stylometry:

We define adversarial stylometry as the notion of applying deception to writing style to affect the outcome of stylometric analysis. This new problem space in the field of stylometry leads to new questions such as what happens when authorship recognition is applied to deceptive writing? Can effective privacy-preserving countermeasures to stylometry be developed? What are the implications of looking at stylometry in an adversarial context?

Brennan et al. (2012) considers adversarial stylometry to be an essential area of study, particularly in terms of privacy and security. Current systems that encourage anonymity in the

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

online environment are primarily concerned with the privacy of an author's location settings, but pay little attention to the privacy of the data itself (Rao & Rohatgi, 2000). The concept is that computer and mobile phone users have a fundamental right to online privacy and that there should be mechanisms to secure their identities. Brennan et al. (2012) argue that present circumvention technologies and the security and privacy community as a whole do not address writing style as a sign of identity. Given the remarkable accuracy of even the most basic stylometry systems, this problem cannot be ignored.

Evidently, antagonistic stylometric research aims to equip individuals with the knowledge necessary to avoid bad online behaviours such as cyberbullying and retaliation. This research is especially valued and used by journalists, corporations, activists and law enforcement officials (Brennan et al., 2012; Kacmarcik & Gamon, 2006). While it is true that individuals who publish messages or other kinds of text on the internet have a right to privacy, it must also be considered that some individuals may use adversarial stylometry techniques to conceal their identity during illicit acts.

The research indicates that the author of a text might camouflage his or her writing style using three strategies:

- The author of a text may initially make a concerted effort to disguise his or her own writing style. This method is known as darkening.
- Second, the author can attempt to replicate the writing style of another author by imitation. This means that author A examines the works of author B and then develops a manuscript that resembles author B's language usage and writing style as closely as possible, or at least contains a substantial quantity of these traits.
- Text translation by machine is the third technique an author may employ. A translation application enables the author to alter the word order and word selection within a text. The author first translates the content into a different language, such as German, and then back into the original language. This process will result in some words having been changed. In addition, the author may translate the content into two languages before retranslating it back into the original language. Brennan et al. (2012) observed that the latter technique is extremely problematic in particular language groups, owing to the fact that the sentence structure of the translated text sometimes changes, resulting in a text that is no longer coherent.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

Moreover, testing adversarial stylometric approaches is complicated by the fact that some people are better than others at concealing their writing style, so that the efficiency of adversarial stylometric methods may depend more on the individual than the method. Brennan et al. (2012) advise developing a generic writing style to address this issue. The Anonymouth programme is a potential answer to the problem for authors; however, it is still in the prototype phase of its development and requires refinements and tweaks to perform efficiently in a range of settings (Brennan et al., 2012). Anonymouth facilitates the process of disguising an author's writing style by recommending potential adjustments to the writing. These recommendations are based on studies of current stylometric techniques (Brennan et al., 2012; Afroz et al., 2014). In addition, Kacmarcik and Gamon (2006) believe that there should be a generic strategy for concealing an author's writing style. In their statistically based research, it is stated that anonymity may be attained by making fourteen alterations per thousand words. Moreover, Kacmarcik and Gamon (2006) assert that the unmasking technique employed by Koppel and Schler (2004) to validate the authorship of a particular text can be adapted for use as an adversarial stylometric technique. In other words, upon determining (by unmasking) which of the author's attributes reveal him or her as the author of a work, the author might adjust or adapt those properties to conceal his or her writing style.

It is evident from the preceding discussion that adversarial stylometry can easily circumvent contemporary stylometric methods in certain circumstances. For this reason, Brennan et al. (2012, p. 21) propose that stylometric methods be tested in 'worst case' scenarios. These might be situations where the principles of adversarial stylometry have been applied to disguise writing. The fact that stylometry may be used by authors to enhance their concealment of authorship shows that there is a need to evaluate the resistance of stylometry methods to adversaries in scenarios where such practices are likely.

Rao and Rohatgi (2000) assert that the principal component analysis that forms the basis of their research is highly effective in an adversarial setting if used in conjunction with misspelling lists and certain classical stylometry measures, and an assessment of the frequency of function words in a text. Kacmarcik and Gamon (2006) believe that the unmasking method used by Koppel and Schler (2004) is the safest stylometric method to apply in an adversarial context, as it is quite resistant to concealing an author's writing style.

2.6.2 Shorter texts and stylometry

Very short texts are troublesome for stylometric analysis because they contain insufficient linguistic data to be processed (Barry & Luna, 2012). Ideally, the linguist needs a substantial amount of text in order to be successful with a stylometric study. Stamatatos et al. (2001) explain that the fact that most stylometric studies are designed to evaluate lengthy literary texts is one of the reasons why some stylometric analyses fail in shorter texts. Moreover, Stamatatos et al. (2012, pp 196, 208) contend that text lengths of less than 1,000 words are unsuitable for stylometric analyses that focus on the lexical features of an author's language use:

It appears that a text length of less than 1 000 words is insufficient to adequately capture the characteristics of an author's idiosyncratic style using lexical measures, the presented collection of style markers, or a combination of them.

In other words, for an analysis of shorter texts to be useful, a stylometric technique needs to be employed that is designed to produce correct results with shorter texts. This enquiry would aim to construct a forensic-linguistic scenario that is as realistic as feasible. The linguist attempting to determine the authorship of phone text messages may not have access to a large number of texts; and even if they do, the length of the messages may pose a problem, as a huge number of text messages are required to equal a full-length text (i.e., about 1,000 words in length). Chaski (2001, p. 4) argues that shorter texts and limited data are typical in forensic linguistics: '[...] forensically significant records are frequently brief and cannot be expanded; in fact, even well-known materials are frequently brief and limited in amount.'

In some instances, it is nevertheless possible to determine the authors of a questionable text, despite the lack of data available in various author identification scenarios. First, as Morton (1978) states, collocations of inconsistent terms and incorrectly used words ('of the', 'in the', 'along with') may be used as discriminators. Hubbard (1995) cites the above-mentioned study. Second, word pairings can also serve as a discriminator; and third, the frequency with which specific words are used is also seen as an extremely important discriminator. Hubbard (1995) notes that he used the third of these discriminators in his research, since the texts he examined were brief and he lacked adequate text to incorporate sufficient collocations and word pairs. Hubbard (1995) truncated the texts to be compared to a specific word count so that all the texts were the same length as the shortest text (783 words). This was because Hubbard wished to determine the accuracy of a stylometric analysis on texts of roughly the same length as the

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

suspect text. Hubbard used the Chi-square test to compare and contrast the frequency of words across different texts and within the same text. According to Hubbard (1995), this method achieved sufficient success to be considered reliable in the context of his study. However, he states that the results of stylometric analyses on shorter texts should be interpreted with caution, as the content of shorter texts in other contexts may lead to unreliable results.

Based on the preceding discussion, it appears necessary to determine whether it is possible to increase the reliability of the results obtained from a small corpus by focusing on aspects such as spelling, function words, the inclusion of English words, logograms or symbols and the use of punctuation, in addition to the frequency of certain words in text messages. These features are not necessarily under the author's conscious control, and as a result, it is feasible that they might give rise to distinctive writing styles in text messages. As previously noted, idiolect is a very complex notion, and it is not always easy to identify idiosyncratic writing styles or provide evidence for their existence. Idiolect and its associated difficulties are discussed in the next section.

2.7 Idiolect

It is evident that author identification in texts, particularly in short texts/documents, is not a simple operation. Several concerns with author identification, such as the length of texts, have been discussed in this research; however, the most disputed component, idiolect, is worthy of further investigation and discussion. Generic language use is a notion used in the literature when discussing the concept of idiolect. Therefore, it is vital to examine this first.

2.7.1 What is generic language use?

The term 'generic' refers to something that is applicable to a group as opposed to an individual. The assumption that generic language use exists opens up the possibility that departures from generic language exist, which would constitute the idiolect of a particular language user. Generic language use, as observed in written texts, refers to qualities such as spelling, sentence structures, and abbreviations of words based on defined conventions within a given language. Most language users of a certain language adhere to these conventions. To some extent, though, the principles may be bent to create an idiosyncratic use of language. In other words, generic language usage acts as a standard against which idiolectic language usage can be measured. However, defining generic language use and idiolectic language use is complicated by the fact

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

that characteristics of language usage that were initially idiolectic become so widespread that over time they become absorbed into the generic form of the language. Further problems with idiolect are discussed in Section 2.7.2.

2.7.2 What is idiolect?

The term idiolect refers to a person's particular, distinctive use of language, created subconsciously. Individuals are typically unaware of the idiolectic terms and idiolectic methods they employ when speaking or writing. There are numerous definitions of idiolect; four are referred to below. All of them are based on the core concepts of individuality and uniqueness:

An idiolect (Bloch, 1948: 7) is a variety of language developed by the individual speaker as a uniquely patterned aggregate of linguistic characteristics observed in his or her language use, often called 'individual characteristics' in forensic science (McMenamin, 2010, p. 487).

The linguist approaches the problem of disputed authorship from the theoretical standpoint that every native speaker has their own distinct and individual version of the language they speak and write, their own idiolect, and 'this idiolect will manifest itself through distinctive and idiosyncratic choices in texts' (Coulthard, 2004, p. 431).

According to Turell (2010), idiolect should not be used in forensic circumstances. Turell (2010, p. 217) argues that the term idiolectic style underlines the fact that 'each individual favours specific linguistic characteristics that comprise their individual use of language'. Barber (2004) states, '[I]diolect, if such a thing exists, is a language that can be thoroughly described in terms of the intrinsic features of a single person at a given time, a person whose idiolect it is at that time'.

2.7.3 The problems surrounding idiolect

Idiolect is not a straightforward notion, as the preceding definitions demonstrate. When idiolect or idiolectic style is presumed to exist, the forensic linguist is frequently faced with the following question: Is idiolect always discernible, and is it reliable evidence of an author's identity? The forensic linguist should, ideally, conduct an idiolectic study of the suspect author's normal language usage rather than focusing on uncommon terms that the author may use. This requires the linguist to be able to discern idiolect in texts containing ordinary and everyday language usage. Although rare words can contribute to an author's identification, it

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

is possible that they are absent from manuscripts that have been branded suspect simply because they are uncommon in normal speech (Juola, 2006).

Even if the assertion that an individual idiolect exists can be sustained, Grant (2010) argues that there is no assurance that an individual's idiolect can be identified in all texts. There is constant diversity in the manner in which every individual speaks and writes. The variations that occur are classified as intra-variation (variations in one person's speaking or writing style) and inter-variation (variations between two or more people's speaking or writing styles) (Gavaldà-Ferré, 2012). Inter-variation may exist in the form of variance between texts by different authors with few, and in some cases no, commonalities between the texts' parts. Crankshaw (2012) believes that when group-level variance occurs, individual-level variation will also arise. Crankshaw (2012, p. 3) cites Anshen (1978), who states, '[...] not only do any two members of what ought to be the same speech community utilize various variants of the same linguistic form, but so does each individual member.'

The so-called uniqueness of speech principle explains this variance (Chomsky, 1965; Halliday, 1975). According to this theory, texts written by two distinct individuals on the same topic will differ significantly, as will texts written by the same individual at various times. This is because each individual makes various lexico-grammatical decisions at different times (Crankshaw, 2012). These lexico-grammatical differences result in intra-variation, which can make idiolect identification more difficult.

Despite these challenges, the notion persists that, notwithstanding heterogeneity in peoples' speaking and writing styles, there are specific words or phrases typically used by the individual, and that these words or phrases may be used to identify idiolect. Gavaldà-Ferré (2012, p. 262) cites Rose (2002), who states, 'The larger the ratio of between-speaker to within-speaker variance, the easier it is to identify speakers.' According to Gavaldà-Ferré (2012), intra-variance displays less variation than inter-variation. This indicates that a speaker's or writer's own language use changes over time, but that the newer form retains some commonalities with the earlier or older forms:

[...] the results for the experiments that have been conducted show two main important factors to be considered. On one hand, intra-speaker comparisons give results that are closer to one, and therefore show slow variation. These results confirm the hypothesis that a speaker's 'idiolectal style' seems to remain quite stable despite the course of time and a long-term situation of language contact. On the other hand, the inter-speaker variation has proved to be higher than intra-speaker variation which confirms the proposal formulated in section 1 that although there

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

is intra-speaker variation, each speaker has a unique 'idiolectal style' that separates them from the rest of speakers from the same community (Gavaldà-Ferré (2012, p. 270).

According to Coulthard (2004, p. 431), it is possible to erroneously conclude that it is feasible 'to devise a method of linguistic fingerprinting – that is, that the linguistic impressions created by a given speaker/writer should be usable, like a signature, to identify them'. Coulthard's cautionary note here echoes the findings of Gavaldà-Ferré (2012), who clearly shows that intra-speaker variation makes such 'fingerprinting' an unattainable ideal. Coulthard (2004) stresses the point by stating that the concept of linguistic fingerprinting is a deceptive metaphor, and that idiolect identification cannot be equated with linguistic fingerprinting.

Crankshaw (2012), too, expresses scepticism over the use of the phrase 'linguistic fingerprint' as a synonym for idiolect. According to Crankshaw (2012) and Coulthard (2004), an idiolect is not as singular or as unchanging as a fingerprint. The physical fingerprint uniquely identifies an individual because each sample is exactly the same as all others, and exhaustive, in that it holds all the necessary information for identification. In contrast, even a massive sample of linguistic data can only provide partial information about an individual's idiolect.

It should also be acknowledged that external variables affect and alter a person's language usage. Among these variables may include social class, level of education, age and gender. Crankshaw (2012, p. 3) cites McMenamin (2002) as stating that these elements result in 'small variances in their (the individuals') internalised grammar, which then shows in a person's speech, writing, and interactions with others'.

Although the existence of idiolect or idiolectal style is contested, it is generally accepted that if idiolect exists, it will be simpler to discern if the linguist has access to many texts. Also important is the length of the texts; the more information contained in each document, the easier it is to detect individual language usage. However, forensic linguists rarely receive texts that are longer than 750 words. According to Crankshaw (2012), shorter texts can be analysed, although not as thoroughly as longer ones. This means that constraints are placed on the process, which prevents a thorough, comprehensive depiction of the individual's idiolect.

Grant (2010, p. 509–510) states that the linguist must not only be familiar with idiolectic and idiolectic patterns, but also, and more importantly, be able to methodologically demonstrate that these patterns exist:

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

A theory of idiolect must explain why one person's production is consistent across texts and why that person's language is distinctive relative to that of other people.

Although the qualities of x's idiolect are related stipulative to inherent properties of x, this does not preclude the possibility of two unique individuals sharing an idiolect or having idiolects that overlap greatly.

Evidence of idiolect in a particular text is important, as the view that idiolect is present in a certain text or texts is insufficient to establish that the work was written by a particular author. This implies that idiolect must be demonstrable or recognisable. By comparing texts with other texts and by comparing writings by the same author, the forensic linguist strives to demonstrate idiolectic language usage. According to Grant (2010), idiolect is a challenging topic to convey. Grant alludes to the intricacy of idiolect when he states, 'Consistency and distinctiveness may be evidence that an idiolect exists, but they do not constitute an idiolect explanation theory' (Grant, 2010, p. 509). This means that a text or texts may possess consistency and distinctiveness traits that may be used to identify an author, but the linguist may still struggle to explain why these characteristics are idiolectic of the specific author.

2.7.4 Theories about the presence of idiolect

There are currently three main theoretical perspectives on idiolect. The first is a cognitive theory based on the idea that the linguist must understand the cognitive mechanisms that permit text production. The second approach holds that a stylistic comprehension of language and language production is sufficient to explain written materials' consistency and distinctiveness (Grant, 2010). The third perspective, Systemic-Functional Grammar (SFG), postulates that there are three inherent sorts of diversity in language: *register*, *code* and *dialect*.

2.7.4.1 The cognitive approach

According to cognitivist theory, an individual's linguistic competence determines his or her language production. Grant (2010) defines linguistic competence as the cognitive ability to produce language. This is represented in the use of words. Aspects such as syntactic complexity and the size of an individual's lexicon can, to a limited extent, be determined on the basis of cognitivist theory. These observations and measurements may be used to indicate variations across people and groups to a limited extent. Grant (2010) cites instances in which quantitative and computational linguistics, particularly in longer texts, have been used to numerically explain the features of the language creation of humans (Holmes, 1998; Grant, 2007; Chaski,

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

2001; Grant, 2008). Although cognitivist theory may explain why a person's language production contains patterns and permanent characteristics, it does not provide the necessary explanation for idiolect.

2.7.4.2 *The stylistic approach*

In terms of idiolect-related perspectives and research, stylistic theory is opposed to cognitivist theory. Theorists in the discipline of stylistics argue that stylistics is vital for comprehending individual linguistic distinctions. This position is not accepted by cognitivist theorists, who claim that stylistic theory is not founded on a consistent language theory. However, according to Grant (2010, p. 512), 'Viewing language variation stylistically as the interaction of habit and situation does not imply a lack of linguistic theory so much as an alternative linguistic theory.' Grant (2010) notes further that cognitivist theory and sociolinguistic studies cannot, on their own, determine idiolect. Instead, emphasis should be placed on a unified theory of idiolect in which cognitivist and stylistic theories are merged so that the strengths of each theory may be leveraged. Such a coherent theory could provide more convincing evidence and explanations for the existence of idiolect.

[...] although cognitivist theories can provide convincing explanations for some aspects of language production these theories hold less power in and of themselves in explaining individual variation. Conversely, while stylistic approaches to the linguistic individual do concentrate on providing explanations for language variation between individuals, they are perhaps less interested in explaining how these might be realised psychologically (Grant, 2010, p 214).

The generally used stylometric approaches for author identification (such as determining the frequency of function words, general sentence length, synonym use, and general paragraph and word length) do not mention idiolect explicitly. In other words, the researchers do not label their research as analyses of the idiolects of persons to identify the authors of texts, although this is precisely what is being investigated. Stylometric methods explore distinguishing characteristics of authors' writing styles to identify the author of a specific piece. These characteristics of an author's writing style are precisely characteristics that the author does not actively control. Despite this, the researcher believes that it is a reasonable argument to assert that idiolect exists and plays a significant role in author identification.

The observability of idiolect, however, is a different matter. Although stylometric methods use idiolect to identify writers, it should not be assumed that idiolect is always evident in the writing styles of authors. Evidently, if a forensic linguist has access to a vast number of texts,

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

determining the idiolects of various authors will be made easier. The more material available to the linguist, the easier it will be to demonstrate not just that idiolect is present in an author's writing style, but also how the idiolects of two authors differ from one another.

A focus on function words in a text is an important part of author identification already highlighted by forensic linguists (Mosteller & Wallace, 1964; Holmes, 1998), and may be of relevance in identifying idiolect. Thus the primary emphasis of an enquiry into idiolect may not be the most prominent words used in a text, but rather the use of non-context-specific terms.

Idiolect is the foundation of author identification research. The fact that authors may be distinguished from one another lends credence to the claim that each author has a unique idiolect. However, the flawed accuracy of some author identification methods (which achieve very low percentages of certainty in terms of authors identification) supports the argument that idiolect is a subtle area of enquiry. In legal matters, the forensic linguist must be able to demonstrate with reasonable certainty that idiolect is present, and that one person's idiolect can be distinguished from another's.

2.7.4.3 Systemic-functional grammar (SFG)

Systemic-Functional Grammar (SFG) is defined as 'a descriptive and interpretive framework that can be used to examine language as a strategic, meaning-shaping resource' (Eggins, 2004, p. 2). Eggins (2004) explains further that proponents of this theoretical framework make four statements about language: (i) language is functional, (ii) language's function is meaning generating (semantic), (iii) this meaning is influenced by the social and cultural context in which the act of language occurs, and (iv) language use is a semiotic process, i.e., a process in which meaning is created by exercising choices. From the aforementioned claims, it becomes clear why this theoretical framework may be applied to authorship investigations and why it may be used to reconcile Nini and Grant's (2013) cognitive and stylistic approaches. Points (iii) and (iv) are especially relevant to authorship research and support the claims of the stylistic approach to such research. The variables that are central to SFG highlight the diversity in language and may assist with characterising a person's language preferences and cognitive processes. The recognition of context, choice and diversity makes SFG theory particularly relevant to the cognitive approach to authorship studies.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

According to Matthiessen (2007), the SFG presupposes that there are three sorts of inherent variety in language: register, code and dialect. In this approach, register variation refers mostly to variation that results from contextual circumstances. Code variation and/or dialectical variation refers to factors such as social status, age, geographical origin, etc. that cause language variance but are tied to the language user (Crystal, 2008; Nini & Grant, 2013).

a) Register variation

Register variation, according to the SFG, entails determining the probability that an author will use specific language options in an organised manner based on contextual elements (Nini & Grant, 2013). This includes the use of formal language in the writing of an academic text or the modifications one might make to speech when addressing a young child. The SFG specifies three connected and abstract categories of contextual elements that cause register variation: discourse mode (mode), discourse draw (tenor), and sphere of discourse (field) (Schleppegrell, 2013).

Discourse mode relates to the function of the text (or language) in the context, such as the use of spoken or written language (Eggins, 2004). According to Crystal (2008), mode can be more thoroughly described as medium of discourse and encompasses the choice to use spoken or written language, as well as the format of the text (e.g., newspaper, poem, commentary).

Discourse draw or tenor is the social role of a language user in a particular language act scenario (Eggins, 2004). According to Schleppegrell (2013), discourse relates to interpersonal connections in a specific context and involves expressions of attitude or disposition. It focuses on the degree of standardisation assumed by language users (Crystal, 2008).

Discourse field refers to the subject matter of the language act. Eggins (2004) emphasises that language users make different linguistic choices when discussing, for instance, jogging that when discussing linguistics (Eggins, 2004). Therefore, sphere of discourse is related to discourse content (Crystal, 2008).

b) Dialect variation

According to the SFG, dialectical diversity is associated with how semantic alternatives are actualised. The manner in which these semantic alternatives are implemented becomes the standard among specific groups of individuals who interact. This standard presupposes a socio-

geographical interaction between language users (Matthiessen, 2007). Thus, the diversity lies in the different ways in which individuals may express the same semantic unit. An example would be the past tense of *burn* in English, where a choice may be made between *burnt* and *burned*. The fact that some individuals use 'burnt' and other use 'burned' indicates dialectic variance. Individuals from the same socio-geographic background generally select the same form (such as the past tense of 'burn') to achieve a certain semantic unity (Nini & Grant, 2013).

c) Code variation

Hasan and Cloran (1990) define code as the linguistic style adopted by individuals to suit their understanding of a particular setting. This form of variation is frequently confused with discourse draw because both depend on context. Unlike register variation, however, code variation is the individual's interpretation of the context that is at issue, not just the context itself. Therefore, the diversity results from everyone's unique understanding of the circumstances. This approach is a methodological approach, not a theory.

2.8 Writing-style features used in stylometry

Individuals may be distinguished by their relatively consistent writing styles, according to the study of stylistics or stylometry. A person's writing style is described by the vocabulary employed, the choice of special characters, and the sentence structure they use. Studies demonstrate that there is no one set of optimum and universally applicable qualities (Iqbal et al., 2010). There are four categories of stylometric features: lexical, syntactic, structural and content-specific. In the following section, each of these characteristics is described in detail.

2.8.1 Lexical features

A text may be understood as a series of tokens organised into sentences. A token may consist of a letter, a number or a punctuation mark. Earlier studies of authorship attribution relied on simple metrics such as sentence length and word count. The benefit of these features is that they may be applied to any corpus in any language, with no additional prerequisites other than the presence of a tokeniser (Stamatatos, 2009).

Lexical features are used to determine an individual's preferred use of characters and words. These characteristics include, for example, the frequency of individual letters, the frequency of special characters, the total number of capital letters, the use of capital letters at the beginnings

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

of sentences, the average number of characters per word, and the average number of characters per sentence. Text may also be understood as a string of characters. Stamatatos (2009) states that numerous character-level measurements can be defined, including the number of alphabetic characters, the number of digit characters, the number of uppercase and lowercase characters, letter frequencies, and the number of punctuation marks.

Vocabulary richness quantifies a text's vocabulary diversity. This measure, also known as word-based features, includes the ratio $V:N$ (where V is the size of the vocabulary and N is the total number of tokens in the text), Yule's K -measure, and the amount of hapax legomena (words that occur once) and hapax dislegomena (words that occur twice) (De Vel, 2000). Unfortunately, according to Stamatatos (2009), vocabulary size varies strongly according to the length of the text. Several functions, such as Yule's K -measure, Simpson's D -measure, Sichel's S -measure, Brunet's W -measure, and Honore's R -measure, have been proposed to attain stability across text lengths. Extracting the most frequent terms from the corpus is another way to define a lexical feature collection. Argamon et al. (2005) employs terms that exist at least twice in the corpus.

From a different perspective, Koppel and Schler (2003) presented various writing mistake measurements to capture an author's unique writing style. In order to accomplish this, they established a set of spelling errors (letter omissions and insertions) and formatting problems (all capital letters) and presented a method for extracting these metrics automatically using a spell checker.

2.8.2 Syntactic features

According to Abbasi and Chen (2005), syntactic features are sentence formation patterns, and involve the use of sentence-structuring tools. Included among these are punctuation and function terms, with common function words (articles, prepositions, pronouns) being 'while', 'upon', 'though', 'where', and 'you'. The following are studies on the use of function words: Abbasi and Chen (2005); Argamon, Saric and Stein (2003); Koppel and Schler (2003); and Zhao and Zobel (2007). The number of function words used by authors ranges from 150 to 675.

Grieve (2005: pp. 18, 32) states that Smith's 1888 study was the first to use function words to identify the authors of plagiarised texts. Grieve (2005) identifies Mosteller and Wallace's

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

(1964) investigation of the authorship of the Federalist Papers as one of the most well-known examples of author identification, in which the frequency of function words in the texts was used to differentiate the authors from one another. The Federalist Papers investigation is regarded as the first non-traditional authorship identification study (as opposed to so-called standard authorship identification investigations, based entirely on human competence) since Mosteller and Wallace (1964) employed statistical tests to conduct the investigation (Stamatatos, 2009: 539). According to Stamatatos (2009), research in author identification (of written texts) continued until the late 1990s and grew to encompass a number of methodologies. Stamatatos (2009) characterises these methodologies as attempts to specify the characteristics that may be used for assessing writing style. This area of author identification is known as stylometry.

Despite the fact that several studies have demonstrated the successful use of function words and collocations as markers for authorship identification, there are still questions about the reliability of this method for authorship identification. Holmes (1994:91) cites Oakman (1980), who correctly asserts that the case of the Federalist Papers was the ideal scenario for author identification, since it involved multiple lengthy texts with few possible authors. As Holmes (1994) citing Oakman (1980) points out, in such cases the use of function words will lead to positive results, but the same method may not be effective in other scenarios.

2.8.3 Structural features

It is possible to determine how an individual arranges the structure of his documents by analysing structural elements; for instance, one may examine the organisation of sentences within paragraphs and of paragraphs within texts. De Vel et al. (2001) was the first to propose the use of structural characteristics for email authorship attribution. In addition to the generic structural characteristics, the authors of the study examined special characteristics in e-mails, such as the presence/absence of greetings and parting statements and the position of these opening and closing phrases within the e-mail.

2.8.4 Content-specific features

Content-specific features are used to characterise certain activities, discussion forums or interest groups by a few keywords or terms (Iqbal, 2010). In their study 'A framework for authorship identification of online messages: Writing-style features and techniques', Zheng et

al. (2006) manually examine, analyse, and identify 11 key terms (Obo, sale, windows, software, Microsoft...) as content-specific characteristics for English 'for sale' internet messages.

2.9 Computer-based methods/Classification techniques used in stylometry

According to the literature, two sets of texts are considered for every authorship analysis task. The first one is a set of candidate authors; texts of known authors, referred to as the training corpus. The second corpus consists of materials by unknown authors referred to as the test corpus. Each of these texts in these sets of texts must be attributed to an author candidate (Stamatatos, 2009).

One method of managing them is to combine all the available training materials for each author into one file per author, which may then be used to extract each author's style properties. The most likely author for a suspect text is then estimated based on a distance measure between the suspect text and the style properties identified for each author. Consequently, variations in training texts written by the same author are eliminated (Stamatatos, 2009). In the literature, this method is implemented using probabilistic models and compression models. Probabilistic modelling is a statistical approach that uses the effect of random occurrences or actions to forecast the possibility of future results (Stamatatos, 2009). It is a quantitative modelling method that projects several possible outcomes that might even go beyond what has happened recently. Compression models (Stamatatos, 2009) make use of state-of-the-art (SOTA) deep learning models on edge devices that have low computing power and memory and will not compromise the models' performance in terms of accuracy, precision and recall. This process broadly reduces two things in the model: size and latency. Size reduction makes the model simpler by reducing model parameters, thereby reducing RAM requirements in execution and storage requirements in memory. Latency reduction decreases the time taken by a model to make a prediction or infer a result. Model size and latency often go together, and most techniques reduce both.

To construct an accurate attribution model, another range of methods necessitates the use of several training text samples per author. This indicates that each training text is represented as a distinct instance of authorial style (Stamatatos, 2009). In the literature, the second approach is adopted through the use of machine learning classifiers, the clustering of algorithms and calculation of inter-textual distance.

2.9.1 CUSUM technique (CUMulative SUM)⁷

Stamatatos (2009) considers the CUSUM technique (or QSUM technique), developed in the 1990s, to be the most well-known example of a non-traditional, computer-assisted authorship identification method. Initially, this technique was deemed to be so reliable that its data could be used in court and considered expert testimony. However, subsequent research demonstrated that the CUSUM method was unreliable, and the resulting data was deemed invalid (Holmes & Tweedie, 1995). The lack of an objective evaluation method for author identification in the early 1990s is the primary reason why the CUSUM technique failed after being initially deemed reliable. Stamatatos (2009) believes that limitations in evaluation processes led to the unreliability of method evaluations. Among these limitations were: cumbersome data – some trial analyses comprised entire books, and the style of these texts varied; a handful of potential authors (usually two or three); comparisons between methods were challenging; and during the initial methodological tests and comparisons, insufficient benchmark data was available.

Since the development and proliferation of electronic texts such as email, online forums, text messages, blogs and social networks, author identification techniques have advanced significantly. Information retrieval, machine learning and natural language processing have all expanded because of technological advancement and the vast quantity of texts available electronically today. The decades following the year 2000 are considered the new era of authorship identification, with technology being at the centre of these investigations (Stamatatos, 2009). In other words, author identification has shifted from being a computer-assisted field to being a computer-centric field.

1.9.2 N-gram analysis

After the advent of computer programmes for author identification, the n-gram method of author identification was created. Initially, n-gram analysis was used for text and language classification. The method used in n-gram analysis to classify texts or languages may be adapted to identify an individual's assumed idiolect. N-gram analysis is a popular technique because it can be applied to any language or document, regardless of the structure of the language, provided that the document contains a substantial amount of text. N-gram analysis is based on a calculation and comparison of fractional n-gram profiles, according to Cavnar

⁷ In statistical quality control the CUSUM (or **cumulative sum control chart**) is a sequential analysis technique developed by E. S. Page of the University of Cambridge, in 1954. It is typically used for monitoring change detection.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

and Trenkle (1994). First, the n-gram system is applied to the training data to calculate profiles; this becomes the training set data. Depending on the method of categorisation, these profiles are placed in various categories. Obviously, the categorisation of texts, languages and authors will involve multiple sets of categories. The system then calculates a classification profile for a specific document. It calculates the distance between the document's profile and each of the categories' profiles to determine which document profile and category profile are most similar. These two profiles have the shortest distance between them and, as a result, the greatest number of similarities.

The length of an 'n' is an important aspect of n-gram analysis because it determines how much lexical, thematic and contextual information may be extracted from each text. For instance, the researcher can conduct an analysis using bi-grammes, tri-grammes and quad-grammes. In other words, the assumption underlying n-gram analysis is that text consists solely of a string of characters. Through n-gram analysis, these characters, as well as letter frequencies, numbers and use of punctuation, can be counted. Thus, the length of an 'n' determines the size of the analysed pieces. The longer the n-gram, the greater the ability to establish lexical, thematic and contextual information. Longer n-grams also increase the dimensions of the analysis, as hundreds or even thousands of distinguishing characteristics can be generated. Shorter n-grams enable the determination of sub-word information (Stamatatos, 2009). This refers to any information pertaining to the syllables in the text. According to Cavnar and Trenkle (1994), the various n-grams for the word *text* are as follows:

bi-grams: _T, TE, EX, XT, T_

tri-grams: _TE, TEX, EXT, XT_, T_

quad-grams: _TEX, TEXT, EXT_, TX_, T__

Evidently, William Ralph Bennett (1976) was the first researcher to propose n-gram analysis as a potential author identification technique. Keselj, Fuchun, Cercone and Thomas (2003) discovered that n-gram analysis successfully identified authors when they compared the 1500 to 2000 most frequent 6-grammes or 7-grammes. However, before an investigator draws conclusions on the success of n-gram analysis, Grieve (2005) advises that the following on n-gram analysis be taken into account. While the studies achieved an impressive level of precision, it must be acknowledged that their results are based on a questionable experimental

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

design. Specifically, the sets of possible authors that these researchers considered span such a wide variety of dialects, registers, eras, and topics that it is impossible to predict whether their method would be able to distinguish between authors in a more stylistically and thematically homogenous set of possible authors.

In their research, Clement and Sharp (2003) and Khmelev and Tweedie (2001) used n-gram analysis with varying outcomes. Khmelev and Tweedie (2001) used bi-grammes (2-grammes), creating a Markov model in which every possible succession of two letters included the possibility that the second letter in an identified n-gram always followed the same first letter within the writing style of a specific author. This method identified 45 authors with a success rate of up to 74.4%, but the dataset is too unreliable to draw broad conclusions about the results. To successfully draw general conclusions from n-gram analysis research results, the experimental design of the studies would need to be beyond reproach. This implies, among other things, that the datasets used should not be too small or too diverse, since small, diverse samples would now allow generalisations to be made.

Grieve (2005) believes that n-gram analysis is a good indicator of authorship despite the poor design of some studies, because n-grams are sensitive to style aspects such as words, collocations and punctuation.

2.10 Writing-style features identified for this study

Perhaps the most extensive and comprehensive application of authorship analysis is in the field of literature and published articles. Well-known authorship analysis studies include the disputed Federalist Papers and Shakespeare's works. In these studies, specific author features such as unusual diction, frequency of certain words, choice of rhymes and habits of hyphenation have been used as tests of authorship attribution (Zhang, et al., 2014). These authorial features are examples of stylistic evidence, considered useful in establishing the authorship of a text. It is conjectured that a given author's style comprises a sufficient number of distinctive features or attributes to uniquely identify the author (Tkacukova, 2019). Stylometric features used in early authorship attribution studies were character or word based, and made use of vocabulary richness metrics (e.g., Zipf's word frequency distribution and its variants), word length etc. (Zhang, et al. 2014). However, some of these stylometric features could be generated under the conscious control of the author and, consequently, may be content-dependent and are a function of the document topic, genre epoch, etc.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

Rather than using content-dependent features, the forensic linguist should use features derived from words and/or syntactic patterns, since such features are more likely to be content-independent and thus potentially more useful in discriminating between authors in different contexts (Zheng, et al. 2006). It is thought that syntactic structure is generated dynamically and subconsciously when language is created, as various utterances are subconsciously made during speech. That is, some language patterns or syntactic features may be generated in ways beyond an author's conscious control. An example of such features is short, all-purpose words (referred to as function words) such as 'the', 'if' and 'to', whose frequency or relative frequency of usage is unaffected by the subject matter (Zheng, et al. 2006). Another example of a syntactic feature is punctuation, as punctuation may not always be guided by strict placement rules but will vary from author to author. Carole Chaski (2007) has shown that punctuation placement may be a useful feature for discriminating between authors. Therefore, a combination of syntactic features may be sufficient to uniquely identify an author.

2.10.1 Variables identified by Chaski

Chaski was the first linguist to systematically consider markedness in terms of authorship, and her statistical analysis of syntax in authorship has passed the Daubert challenge in US courts. The Theory of Markedness has been studied for over half a century, with both Jakobson (1963) and Chomsky (1965) being influential researchers who sought 'linguistic universals – rules of structure that applied to all languages' (Olsson, 2008:46). Olsson (2008) explains that markedness refers to linguistically unusual or strange structures and signs that, on some level, are non-standard. Markedness may be observed in any aspect of a text that catches the analyst's attention and strikes him or her as odd. The fundamental premise of linguistic markedness is that one identifies a set of specific notable aspects of the text in question, used by the author at various levels of structural preference (Battistella, 1995) or phonological representation, which characterise the segments of language (Kean, 1975). The difference between markedness and idiolect may be summed up as follows: Markedness refers to language choices that are non-standard, while idiolect refers to every person's unique language use.

Chaski formulated a systematic method for authorship identification by designing the programme ALIAS (Automated Linguistics Identification and Assessment System) that implements a syntactic analysis method for authorship identification (Chaski, 2007). ALIAS

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

uses only three types of variables: end-of-sentence punctuation, internal structure of sentences and average sentence length (Chaski, 2007), producing numeric outputs for each and analysing them statistically. The programme was used successfully by Chaski on a homogeneous English-speaking group. The current research did not use the computer programme ALIAS for analytical purposes, mostly because it is not freely available and the researcher has never used it; however, the research did make use of the idea that the above three variables enable authorship identification.

2.10.2 T-units

A T-unit is a measurement in linguistics referring to a main clause plus any subordinate clauses that may be attached to it. As defined by Hunt (1965), the T-unit, or the minimal terminal language unit, measures the smallest word group that may be considered a grammatical sentence. Research suggests that the length of a T-unit can be used as an index of syntactic complexity. It is therefore also classified as a syntactic feature.

T-unit analysis, developed by Hunt (1965), has been extensively used to measure overall syntactic complexity in both speech and writing samples (Gaies, 1980). Hunt (1965) claims that the length of the average T-unit in the writings of an individual mirrors that individual's cognitive development and gives a satisfactory and stable index of language development. The T-unit's popularity is due to the fact that it is a global measure of language development that occurs outside a given set of data and allows meaningful comparison between first- and second-language acquisition (Hirano, 1989). T-unit analysis was successfully used by Larsen-Freeman and Strom (1977) and Perkins (1980) as an objective criterion to evaluate the quality of English in second language (ESL) student writing. Analysis of error and content variables yielded results, as in the case of other researchers who reported that high-rated essays tend to be longer, contain larger T-units and clauses, make use of more non-restrictive modifiers and have fewer errors than low-rated essays (Hirano, 1989). Better writers seem to have higher creative skills, which allows them to expand and develop concepts in writing. It can be helpful for authorship attribution if people are found to use T-units differently, according to their proficiency in English.

2.10.3 Cohesion markers

The linguistic patterns in a text include a reader's experience of reality, and at the same time also provide structure to it; the same patterns also make it possible 'to identify what features of the environment are relevant to linguistic behaviour and so form part of the context of the situation' (Halliday & Hasan, 1976, p. 20). A text is coherent with respect to two aspects: the textual context (i.e., the uniformity in its register), and the text itself (i.e., the cohesiveness of the text). Text and context therefore interact continuously because a reader wants to bring about this interaction (Wybenga, 1988). These two aspects, text and context, by which cohesion and coherence are established in a text, are therefore distinguishable but not separable.

The concept of cohesion in writing was established by Halliday and Hasan (1976) and is considered one of the ways in which the 'texture' of a text is created. Cohesion involves an investigation of the binding patterns in a text created through syntactic, semantic, morphological and phonological methods. Thus, for example, the reference structure in a text can be determined – connections between nouns (noun-noun, noun-pronoun), and between nouns and adverbs, among others, may be determined; in addition, the linguist may look for coherent sound patterns and lexical-semantic linkages. Cohesion chains comprise lexical cohesion, reference, substitution and conjunction (Carstens & Van de Poel, 2012).

Coherence refers to that which contributes to a text making sense and exhibiting coherence (Carstens, 2016). The linguist needs to ask whether there is a conceptual link between what the writer knows and what the reader may infer from the text. The following paragraphs give an overview of the cohesion markers used in writing, showing how each one may contribute to meaning and ultimately to the cohesion of the text.

2.10.3.1 Reference

The traditional view of reference, according to Brown and Yule (1983), includes the following two aspects:

- (a) the relationship between expressions and the text and between the text and entities in the extra-lingual world; and
- (b) coreferences between expressions in different parts of a sentence or text.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

Reference therefore forms an important part of the written communication process. Certain language elements (words and expressions) establish connections between the ideas inherent in units of text; between the text and extra-lingual reality; and between elements in the same piece of writing (Carstens, 2016). Reference is the most prevalent chain of cohesion in language use. It rests on the principle that two (or more) elements in the same passage may be semantically associated with each other because both refer to the same referent in reality.

De Stadler (1989) identifies four main types of reference:

- i. Particular reference: *George's bike* is outside; *he* didn't put *it* away last night.
- ii. No particular reference: I saw *a man* with her.
- iii. Generic reference: *This vehicle* is a wonderful mode of transport.
- iv. Unique reference: *Kelly Johnson* drives the newest car on the market, the *Audi TT*.

2.10.3.2 Substitution

As a cohesion marker, substitution in formal written language is fairly rare. Substitution refers to the *replacement* of one element by another element (a substitute word or expression) in such a way that the meaning of the sentence is not affected (Carstens & Van de Poel, 2012). It is striking, however, that the element replacing the original word does not refer to the same referent as the antecedent (in the majority of cases) (Carstens & Van de Poel, 2012).

The definition of substitution given by Carstens (2016) is: One element in a sentence is replaced under certain (grammatical and contextual) circumstances by another (exchange) element that can act in the relevant element's place. The following is an example:

My pencil is too short to write with anymore. I will have to buy *a new one*.

It is clear here that 'a new one' does not refer to the exact pencil which has become too short to write with.

2.10.3.3 Ellipsis

The use of ellipsis (plural ellipses) is common in speech and even more so in written language. This phenomenon (also referred to in the literature as contraction, continuance or deletion) involves the omitting of certain elements from a passage because the omitted parts within the context of the language expression may be inferred (Carstens & Van de Poel, 2012). In the case

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

of substitution, the part being replaced is still there, but in the case of ellipsis the word's presence must be inferred. The literature refers to this as 'substitution by zero'. Examples are

- The policeman moves slowly towards the sleeping man and [the policeman] shakes his shoulder.
- The attacker hit him over the head with a rock and [the attacker] robbed him of his wallet and [the attacker robbed him] [of his] car keys.

2.10.3.4 Conjunction

Conjunction as a phenomenon is particularly common in texts – conjunction helps to make a text 'flow'. A text making correct use of conjunctions exhibits coherence, in that the words preceding and following a conjunction are linked by means of the appropriate conjunction. The conjunction or conjunction marker indicates the nature of the relationship between the two parts of the sentence (Carstens & Van de Poel, 2012). The interrelationships expressed by conjunctions include contrast, reasoning, time indication, summation, ordering, analysis, illustration, interruption and conformity (Carstens & Van de Poel, 2012). Conjunctions and adverbs typically act in language as markers of conjunction. Consider the following examples:

- Mike is intelligent, hardworking *and* honest. *In short*, he is the right person for the job. (Linking)
- He hit the wall with his fist *because* he was angry. (Reason)
- We had been in the queue for a long time to buy tickets, *but* we *finally* got them. (but = contrast; finally = time)

2.10.3.5 Lexical cohesion

Lexical cohesion concerns linkages between so-called content words (nouns, verbs, adjectives and adverbs) used in successive sentences (Carstens & Van de Poel, 2012). Two kinds of lexical relationships (i.e., meaning relationships) are indicated by these words: repetition and collocation (Carstens & Van de Poel, 2012). Repetition is what it says; the repetitions of words. Collocation refers to the use of words that are associated with each other at the meaning level, even if the nature of the relationship in question is not entirely clear. In such cases, the evident relationship between certain words contributes to the bonding between them, which enhances the cohesion and coherence of the text.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

The role of lexical repetition is relevant here – in particular, the types of repetition that may be discerned and the functions of these forms in language texts.

Lexical repetition occurs where one lexical item refers to another, related item by means of a referent that they have in common (Stotsky, 1983). This reference to a previous common referent may be done by direct repetition of the lexical item or by reappointment of the meaning by exploiting the lexical relations (Carstens, 2016). Stable semantic relations exist between words and are the basis for the descriptions given in dictionaries. Hoey (1991) adds to this by stating that repetition covers a whole range of ways in which one lexical item may be understood or may evoke the essence of a previous lexical item. Halliday and Hasan (1976) distinguish the following types of lexical repetition: direct repetition, synonyms, superordinates and common nouns (epithet).

- Direct repetition of the same lexical item

The most obvious form of lexical repetition is the use of the same word later in the sentence or in later sentences.

- Synonyms

Synonyms may be used to indicate a lexical relationship to a foregoing lexical item. The popular view of synonyms is that they are words that have the same or approximately the same meaning. The more semantically based view is that synonyms are words that have one or more meaning distinctions.

- Superordinates

The overarching concept in a lexical group can also act as the linking element. For example, ‘Please send me *roses*. These are my favourite *flowers*.’ The overarching concept of ‘flowers’ serves to continue the reference to the hyponym ‘roses’.

- Epithet

A common word may also serve as a repetition of the original element. For example: ‘Look how my *grandmother* kneads that dough! This *old woman* can really bake.’

Lexical cohesion may make use of the following lexico-semantic relations, among others: association with a particular topic or theme, opposition, contrast, membership in an ordered lexical series (e.g., days of the week, month names), and membership in an unordered series (e.g., colour terms, automobile terms, cake terms, economy terms, and so on). The words

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

involved in these relations stand in some semantic relation to each other, but for textual linking purposes the exact nature of the relationship may not always be relevant; what matters is that the words can be associated with each other in some way.

Collocation refers to the relations between lexical items in a semantic field, such as synonyms, antonyms, hyponyms, part-whole relationships, ordered series, and unordered series (De Stadler, 1989). Syntagmatic relationships are also relevant here.

- Synonyms

As in the case of lexical repetition, synonyms used to indicate collocation denote a connection between terms, but here the connection rests on a concordance in meaning, or, as De Klerk (1978, p. 103) says, on ‘... the relation between two or more words that have the same connotation’. The context-boundedness of synonyms plays a cardinal role here. Compare the following series: ‘story – fairy tale – fable’, ‘cunning – sly – sneaky’ and ‘beautiful – pretty – gorgeous’. Each group of three words is an example of collocation.

- Meaning opposition and contrast

Different forms of meaning opposition may be used; namely, antonyms, complementarity and inverses.

An antonym is a word that expresses an opposite value to another within a certain context. The contrast is gradable. The gradability of these forms can be expressed grammatically, among other things, in the three steps of comparison.

Complementary terms do not exhibit the same degree ability as antonyms, nor do they take steps of comparison. The affirmation of one term involves the negation of the other complementary term, and vice versa (e.g., *true* and *false*).

Inverses also do not allow grading, and the negation/affirmation of one term does not necessarily entail the negation/affirmation of the other term (e.g., *father* and *son*, *get* and *received* and *go* and *come*).

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

De Stadler (1989: 87) also mentions 'idiomatic opposites' as a form of meaning contrast (e.g., *tea and coffee, knife and fork and salt and pepper*).

- Hyponyms

Hyponyms refer to words of more specific meaning than the general term (the superordinate), applicable to it; for example, *rose, dahlia, carnation, tulip and protea* are hyponyms of the superordinate *flower*.

- Part-whole relationships

These kinds of relationships also play an important role in collocations. Part-whole relationships are expressed in hyponymic word pairs in which the first member of each pair is in a 'part of' relationship to the second member (De Stadler, 1989); for example, *leg and body, string and guitar, saddle and bicycle*, and so on.

- Ordered and unordered series

Words can also occur in ordered and unordered sequences, where the connections between the words in a sequence is their focus on a particular issue or topic. For example, the time terms *second, minute, hour, day, week, month, year, decade, century* occur in an ordered sequence (in other words, one follows the other), while the colour terms *red, yellow, green, purple, blue* are an unordered series (i.e., the order of the words makes no difference to the meaning of the sentence).

- Syntagmatic relations

Syntagmatic relations have to do with the combination possibilities of lexical items with each other. De Stadler (1989) points out that the syntagmatic relations are apparent in the predictable connections that various lexical items have with other lexical items; they connect on a syntagmatic level. For example, *dog and bark, fish and water, drive and car*.

The above discussion of cohesion markers provides a good indication of the areas in which small differences may occur between authors' texts. Cohesion markers are a subtle feature of language. Writers who are aware of the principles of good writing will use cohesion markers differently than people who 'accidentally' use these words. Consequently, these areas are worth look into for evidence of authorship.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

In this study, the choices made by eight authors with regard to the above categories of cohesion markers, and the existence of patterns in their position and frequency of use, were examined with a view to contributing to the field of authorship identification. The study focussed on the position, frequency and patterns of use of five variables specifically: end-of-sentence punctuation, internal structures of sentences, average sentence length, T-units and cohesion markers.

2.11 Conclusion

This chapter has presented a comprehensive review of the field of forensic linguistics and its sub-categories. The origin of the field and the emergence of authorship analysis in South Africa was described, researchers in the field were identified, and the general lack of authorship analysis studies in South Africa was highlighted. The discussion also covered the forensic linguist's role in the courtroom and the ways in which courts in various countries handle linguistic evidence. Evidently, all nations share a need for a dependable and quantitative method for authorship analysis.

The chapter has also proved and substantiated the need for a method of author identification that is acceptable to the courts and satisfies the external requirements for expert evidence as imposed by the judiciary. This need was identified in Section 1.6, 'Objectives of the study'. The chapter has shown that, since there is no universally accepted method, the research carried out for this study could be of great significance to the practical applicability of forensic linguistics in the judicial system.

Today, the term stylistic or stylometric analysis is used to refer to nearly all author identification techniques. The forensic linguist may identify a variety of features in a single text and conduct statistical tests to determine whether their hypothesis about authorship is correct, making use of advanced computer-based statistical software. It is common for forensic linguists to examine various stylistic and stylometric characteristics of the text based on their knowledge of theoretical linguistics, in order to make statistically tested hypotheses about the text or the author.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

According to Sierra et al. (2013), stylometry includes ‘word and character n-grams, punctuation, function words, vocabulary richness, part of speech frequencies, word collocations, grammatical errors, and word, sentence, and paragraph lengths’. In addition, these authors define stylometry as ‘the study of an author’s style through the identification of his or her stylistic characteristics’. In this investigation, both stylistic and stylometric analyses were conducted.

It is worth noting that while great advances have been made in the techniques of author identification, the three main scenarios with which forensic linguists are concerned have not changed significantly. This has allowed forensic linguists to concentrate their research on finding solutions to specific problems.

In the first scenario, the linguist must be able to determine whether a single author is responsible for all the texts in a particular collection. In other words, the linguist must determine whether a text matches other texts by the same author.

In the second scenario, there is more than one suspect author, and the task is quite different; the linguist is required to compare a suspect text with the texts of a group of possible authors. In cases where the suspect has already been identified as a possible author by methods outside of linguistics, the forensic linguist must be able to associate a text with the suspect (McMenamin, 2002).

The third scenario is the most typical of the three. More often than not, when the police approach a forensic linguist, a suspect has already been identified. The police then enlist the assistance of the forensic linguist to prepare an evidentiary case against the suspect (Grant, 2010: 514).

According to Sierra et al. (2013), a fourth scenario may arise during forensic-linguistic investigations. In this scenario, the only suspects are a collection of texts. The forensic linguist must use text properties such as language use and style to obtain information about the texts’ author, endeavouring to create a profile of the possible author. Like Sierra et al. (2013), Olsson (2004), considers ‘profile compilation’ as a fourth scenario in author identification.

In the current study, the five writing-style features (end-of-sentence punctuation, internal structures of sentences, average sentence length, T-units and cohesion markers) enabled author

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

profiling of L1 and L2 speakers of English, who were chosen on the basis of the theoretical approaches set out in Chapter Two. It is the opinion of this researcher that the identified features occur subconsciously in the texts of writers and depend on the writer/speaker's proficiency of a language.

Chapter Three presents the methodology used in the study to analyse the eight selected texts.



CHAPTER 3: METHODOLOGY AND ANALYSIS OF DATA

3.1 Introduction

The literature review presented in Chapter Two outlined the broad theories and frames of thought of the relevant fields of study. It was found that the various theories referred to in the field of authorship identification are widely accepted, but that the methodological approaches within the field are still under dispute. It is therefore widely accepted that more empirical research is needed to test and refine theories and methodologies. The purpose of this study is to determine whether patterns in the use of the identified features are peculiar to a specific author and if so, to what extent.

As both qualitative and quantitative methods were used to analyse the data, this study was a mixed method one (Dornyei, 2007; Angouri, 2010). The study makes use of a corpus of texts in order to achieve the research objectives. The chapter begins with a brief overview of the research approach – both mixed methods and the corpus-based approach – after which corpus selection, the analysis process, and the statistical processing methods are discussed.

3.2 Research approach

According to Dornyei (2007), researchers use the mixed method for three reasons, two of which are relevant for this study: First, a mixed method enhances understanding of complex issues and, second, the method makes the study's results more reliable and verifiable (Dornyei, 2007). Angouri (2010, p. 29–30) argues that it is advantageous to use mixed methods in social science and humanities research, citing, among others, Greene (1989), who argues that 'combining the two paradigms (i.e., quantitative and qualitative methods) is advantageous for constructing comprehensive accounts and providing answers to a wider range of research questions'. Dornyei (2007, p. 166) cautions, however, that the mixed method should not be viewed as an 'anything goes' approach, stating that the researcher must ensure that the research methodology and interpretation of data are consistent.

3.2.1 A corpus-based approach: A brief overview

A corpus-based approach is used to answer specific research questions. Although there are a wide variety of linguistic studies where a corpus-based approach could achieve the research objectives, it is important to note that this is not the only approach to reach these specific

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

research objectives. The characteristics and application possibilities of this approach (discussed in Sections 4.2 and 4.3), together with the requirements of the research questions, determine whether or not this methodological framework is appropriate for a particular study.

Biber, Conrad and Reppen (1998) list several characteristics of a corpus-based approach. First, it is empirical in nature, allowing for an analysis of patterns in natural texts. Empirical research is based on the premise that knowledge may be acquired through experience and observation of a phenomenon (Babbie & Mouton, 1998). In this study, observations are made with regard to texts, with patterns in writing style sought in accordance with the research objectives.

Second, the corpus-based approach is used when the researcher wishes to analyse texts created in a natural way (Biber et al., 1998). Texts created in a natural way are those created by a language user and not by a researcher in order to illustrate a theory. Apart from the use of naturally produced examples, a corpus-based approach is relevant when there is a large number of these texts that need to be investigated. Using a corpus-based approach, the researcher identifies and then analyses patterns, regularities and irregularities in texts.

Third, the corpus-based approach makes use of computers and purpose-designed software to search and group specific phenomena (Biber et al., 1998). The software provides automated and interactive possibilities for extracting and analysing linguistic data from the dataset.

This automatic extraction of data leads to the last character trait Biber discusses; namely, that this approach relies on both qualitative and quantitative methods to analyse data and draw conclusions or inferences (Biber et al., 1998). Qualitative data analysis is used to mark or annotate and count phenomena, after which the researcher examines these phenomena and tries to spot underlying patterns. Biber et al. (1998) makes this clear, stating that corpus-based analysis does not only involve counting linguistic uniqueness, but that qualitative, functional interpretations are an indispensable part of the proper use of the corpus.

McEnery, Xiao and Tono (2006) list four outcomes that cannot be achieved with a corpus-based approach. First, a corpus-based analysis of data cannot give an indication of what is possible and impossible in terms of linguistic constructs. Since this type of study tries to record what appears in texts that have come into existence naturally, it can report only on phenomena that appear in the texts, and can say nothing about their normative validity. Second, a corpus-based analysis cannot address the explanatory aspect of the investigation; that is, corpora

cannot be used to determine why a text adopts a specific structure. Thus, it can produce results but not give reasons for possible linguistic patterns or variations. A third limitation mentioned by McEnery et al. (2006) is that the choice to undertake a corpus-based study entails certain types of methodological solutions and must therefore fit the research objectives. Although this is not a limitation peculiar to a corpus-based approach – all methodological approaches have shortcomings – they believe that it is important to formulate research questions in such a way that it is possible to answer them by following a corpus-based method. Finally, McEnery et al. (2006) also mention that researchers should keep in mind that inferences made following a corpus-based study apply only to the specific corpus. They explain that it is possible to generalise according to a representative corpus, but that researchers must guard against unfounded generalisations.

The theory of corpus-based research provides a framework that creates data processing possibilities for studies that investigate linguistic patterns and their extralinguistic causes/consequences (McEnery et al., 2006). This framework may be used in practice in several ways, as is evident in the great variety of research questions and variables that researchers have used over the years making use of this approach. The corpus-based approach has been used in studies on grammar patterns, register/code switching, language diversity, translation, discourse analysis, stylistics and forensic linguistics (McEnery et al., 2006). Because the current study focused on specific language users' writing styles within two chosen genres, it has strong touchpoints with forensic linguistics, stylistics, register/code switching and genre analysis.

3.2.2 Types of research conducted using a corpus-based approach

In their standard work, *Corpus linguistics: Investigating language structure and use*, Biber et al. explain (1998) that linguistic studies traditionally focus on one of two areas: linguistic structure (morphemes, words, sentence construction) or the way language is used. The focus of Biber et al.'s work, and also the focus of this study, is how language is used; in other words, how language users use (and sometimes abuse) the linguistic resources at their disposal.

According to Biber et al. (1998), studies on language use usually have one of two main objectives. First, they may aim to find a pattern in language use and to determine, describe and/or evaluate the uniqueness, clarity or saliency of the pattern. Second, they may investigate extralinguistic factors that influence variations in language use. In this study, an attempt was

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

made to identify linguistic patterns created through the use of cohesion markers and to determine how clearly these patterns emerged when extralingual factors were used as variables.

3.3 Research questions

As a mixed method study, this research made use of both qualitative and quantitative methods to answer to the following research questions:

- i. To what extent is a standardised method for authorship identification/attribution possible, using a certain combination of writing-style features?
- ii. What are the available writing style features, and to what extent are they effective for authorship identification/attribution of L1 and L2 English texts?
- iii. Which classification techniques are effective for authorship identification/attribution of L1 and L2 English texts?
- iv. Can the proposed standardised method be deployed on online texts?

3.4 Research design

In his chapter on research methodologies in the *The handbook of applied linguistics* (Brown, 2004), Brown describes the diverse nature of linguistic research, stating that the traditional description of research, as for example either qualitative or quantitative, or synchronous or diachronic, is insufficient to describe linguistic research. He further explains that these traditional descriptions, although in principle accurate, should be seen as existing on a continuum between qualitative and quantitative research. He lists twelve so-called character traits of research, although these may be more accurately described as elements or variables in the research design. He says each of these characteristics also exists on a continuum between qualitative and quantitative characteristics. He explains all applied linguistic studies may be placed somewhere on the continuum, depending on their use of these elements, maintaining that this approach more accurately describes the research design than the traditional descriptions (Brown, 2004). In terms of this view, this study may be classified as a mixed method study, since the characteristics or elements of the study lie in the middle of the continuum.

3.4.1 Nature of reasoning

From the literature review it became clear that certain aspects of the theories used in authorship studies are still vague or have not yet been proven by empirical studies. This study therefore uses the theoretical frameworks of authorship studies to investigate their validity in specific circumstances. Based on the theories, certain assumptions were made about authorship studies, and, based on these assumptions, the research questions were developed. Mouton and Marais (1990) explain that an inductive approach works in such a way. They stipulate that the researcher in such studies makes use of ‘... general hypotheses or conjectures that guide the research broadly’ (Mouton & Marais, 1990, p. 105). The assumptions that guide this study logically follow each other. The assumptions and hypotheses that guide this study are the following:

- i. A person’s writing style is identifiable based on unique writing-style features;
- ii. A wide range of syntactic features can characterise a particular style of writing;
- iii. End-of-sentence punctuation, internal structures of sentences, average sentence length, T-units and cohesion markers are among the features that characterise a writing style;
- iv. Each author may use end-of-sentence punctuation, internal structures of sentences, average sentence length, T-units and cohesion markers in their unique way;
- v. An author’s writing style is recognisable based on their use of end-of-sentence punctuation, internal structures of sentences, average sentence length, T-units and cohesion markers.

In terms of Brown’s (2008) explanation of the nature of a research design in linguistic studies, an inductive approach, such as the one followed in this study, may be said to lie on the qualitative end of the continuum.

3.4.2 Type of data

The data used in this study is qualitative, since it comprises texts; specifically, texts that are considered linguistic evidence and form part of the corpus for analysis. McEnery et al. (2006) explain that a corpus must be designed to represent a specific language or language variation, and that there are broadly two types of corpora; general and specialised. A common corpus is designed to ‘... serve as a basis for a description of a language or language variety as a whole’ (McEnery, 2006, p. 15). In this study, a specialised corpus was created through the selection

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

of eight texts. These texts were chosen because they were specifically by English L1 and L2 authors – with many texts it is not clear whether the author is a first language or second language speaker. Evidentiary texts are also not easy to come by or released, but the ones used for this study were.

Even though the data is qualitative in nature, the research methodology aimed at quantifying phenomena in the qualitative data. The raw data of the corpus was encoded to create a set of quantitative metadata that was eventually analysed statistically. The analysis process is described in detail in Section 4.5.

3.5 Data analysis

The analysis of the data in this study relies strongly on the conceptual framework adapted from the literature review. The study draws on the framework created by Zheng, Li, Cheng and Huang (2006), with a view to testing its application and effectiveness in authorship attribution in L1/L2 English texts. In their framework, Zheng, Li, Cheng and Huang (2006) identify four types of writing-style features – lexical, syntactic, structural and content-specific features. They extracted examples of these features from the texts, following which, they used inductive learning algorithms to build feature-based classification models to identify authorship of online messages. The authors conducted experiments on English and Chinese online-newsgroup messages. They compared the discriminating power of the four types of features and of three classification techniques: decision trees, back-propagation neural networks, and support vector machines. The experimental results showed that the proposed approach was able to identify authors of online messages with a satisfactory degree of accuracy of 70 to 95%. All four types of message features contributed to the discrimination between authors of online messages. Support vector machines outperformed the other two classification techniques in their experiments. The high performance they achieved for both the English and Chinese datasets showed the potential of this approach in a multilingual context.

In their studies, Iqbal et al. (2008), El and Kassou (2014) and Grieve (2007) also employed variations of the four types of writing-style features mentioned above. Many studies since Zheng et al. (2006) have used feature sets combining lexical, syntactic and structural features. The current study will be the first, to the best of the researcher's knowledge, to use the specific feature set shown in Figure 3.1 below.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

Phase 1:
Data Collection

Secondary data
L1 and L2 English

Phase 2:
Feature Extraction

Identify and tag writing-
style features

Feature Set:
End-of-sentence
punctuation, internal
structures of sentences,
average sentence length,
T-units, cohesion
markers.

Phase 3:
Method Generation

Test using classification
technique:
WordSmith

Authorship
attribution
method

Phase 4:
**Authorship Identification/
Attribution**

Validated authorship attribution
method

Result of authorship
analysis

Figure 3.1: Conceptual framework for authorship attribution

Adapted from Zheng et al. (2006, p. 384)

Using the conceptual framework as a foundation for the methodology, the following steps were followed:

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

- i. The three language elements (end-of-sentence punctuation, internal structure of sentences and average sentence length), of Chaski's programme, ALIAS, were identified in the texts.
- ii. The T-units were identified.
- iii. The cohesion markers were identified.
- iv. The identified features were quantified to analyse how frequently they occurred and which features most commonly occurred for individual writers.

The steps were employed on multiple L1 and L2 English texts, and results were compared. The researcher then drew conclusions on whether L1 and L2 English speakers make fundamentally different choices regarding these specific writing-style features.

3.5.1 Phase 1: Data collection

The study used readily available data obtained from the internet and books. Texts were chosen based on their availability and suitability for the objectives of the study, and were therefore purposefully sampled as examples of L1 and L2 English texts. Linguistic evidence used by Prof Hilton Hubbard (Hubbard, 1994; 1995) in the 1989 extortion case detailed earlier was used, in the form of four L2 texts; six were received from Prof Hubbard, but only four were used, as an equal number of L1 texts and L2 texts was desired. The four L1 texts chosen were found in various different forensic linguistics handbooks' appendices. All were examples of letters threatening bomb explosions: 'The Army of God' letter, the 'Lampley Hollow' letter, a letter by Luke Jon Helder (a pipe bomber) and a letter sent by Theodore Kaczynski (the Unabomber case detailed earlier). The origins of these texts are explained in Chapter Five. Multiple texts by the same author would have been ideal, since several texts by each author would have revealed when and how often each used the chosen writing-style features.

3.5.2 Phase 2: Feature extraction (also known as tagging)

First, the data was tagged manually with a view to marking the identified writing-style features. This allowed the researcher to observe any unique choices made by the authors.

3.5.3 Phase 3: Method generation

In this study, the computer software *Oxford WordSmith Tools* (Scott, 2021) was used as a research tool. *WordSmith* is computer software that allows for the appearance of specific lexical items or labels in their textual context. It provides possibilities for frequency detection and can make concordances. In this study, the tags discussed were affixed to the original texts so that in *WordSmith* certain groupings and concordances would be made based on the tags and not only the lexical items in the texts. It would not have been possible to make these groupings and concordances computational from the raw data.

For this phase of the analysis, *WordSmith*'s concord function was used. This feature allows the user to see all the appearances of any word, phrase, or tag in the context of the text. The user enters the element as a search phrase and the programme then shows the word phrase in its immediate context. In addition, several search phrases can be entered, and all the search phrases will appear at one time on the same screen.

The programme then indicates the number of times each tagged features occurs. This feature helped to answer the research questions, since it revealed each authors' choices with regard to the writing-style features chosen. Notable differences in the numbers showed whether the features (Question ii) and classification techniques (Question iii) were successful at revealing authorship differences, and whether the chosen features could be used in a possible standardised method for authorship attribution.

3.5.4 Phase 4: Authorship identification/attribution

Once the identified features had been quantified for each author and differences in their context and use had been observed, conclusions were drawn regarding authorship attribution. The results of the individual texts were compared to conclude whether L1 and L2 English speakers make fundamentally different choices with regard to these specific writing-style features.

3.6 Conclusion

In this chapter, the researcher has outlined the method followed to answer the research questions and achieve the study objectives. The study made use of a corpus-based method, which has its particular strengths and limitations, as discussed. The design of the study and the

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

analysis methods used for data processing were discussed in some detail. It was established that the study made use of qualitative data, which was analysed quantitatively, making use of a particular computer software.

In Chapter Four, the findings of the analysis are discussed.



CHAPTER 4: FINDINGS AND DISCUSSION

4.1 Introduction

This chapter describes the analysis of the data set and presents the findings. Following the phases outlined in the conceptual framework presented in Chapter Three, this chapter presents (i) the texts obtained for analysis; (ii) the steps taken for feature extraction; (iii) the frequency of tagged features, shown in table form; (iv) a discussion of the findings; (v) a discussion of the effectiveness of the selected features for authorship attribution. The chapter concludes with a discussion of the potential application of the suggested framework.

4.2 Data presentation

In total, eight texts make up the analysed data set (see Section 3.5.1). Four texts were produced by native English speakers and were obtained from freely available forensic linguistics handbooks' appendices. These texts are examples of letters threatening bomb explosions: 'The Army of God' letter, the 'Lampley Hollow' letter, a letter by Luke Jon Helder (a pipe bomber) and a letter sent by Theodore Kaczynski (the Unabomber). Professor Hilton Hubbard provided the other four texts written by speakers of English as a second language from the 1989 extortion case (see Section 4.2.2). This linguistic evidence comprised six L2 texts, but only four were used in order to have an equal number of L1 and L2 texts. All of the L2 essays were produced by native Slavic speakers. All texts were copied verbatim, so that errors remain as they were in the original.

Table 4.1 shows the size of the L1 dataset and Table 4.2 shows the size of the L2 dataset.

Table 4.1: L1 dataset size

	<i>Words per text</i>	<i>Sentences per text</i>
L1 Text 1	264	16
L1 Text 2	327	31
L1 Text 3	418	37
L1 Text 4	1414	76
Total	2423	160

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

Table 4.2: L2 dataset size

	<i>Words per text</i>	<i>Sentences per text</i>
L2 Text 1	831	39
L2 Text 2	950	36
L2 Text 3	855	57
L2 Text 4	1003	37
Total	3639	169

The total number of words in the dataset was therefore 6062.

4.2.1 First language English speakers' texts

These texts were chosen specifically because they were readily available in books already in the researcher's possession. The full texts were not available on websites or were not accessible. The texts were also chosen because they are definitely written by first language English speakers. Most text or case descriptions do not state whether or not the author was a first language speaker, and this aspect cannot be assumed.

4.2.1.1 The Army of God letter

The Army of God (AOG) is an American Christian terrorist organisation, members of which have perpetrated anti-abortion violence (Altum, 2003). According to the Department of Justice and Department of Homeland Security's joint Terrorism Knowledge Base, the Army of God is an active underground terrorist organisation in the United States (Altum, 2003). In addition to numerous property crimes, the group has committed acts of kidnapping, attempted murder and murder. The following threatening letter was sent to law officials.

THE BOMBING'S IN SANDY SPRING'S AND MIDTOWN WHERE CARRIED OUT BY THE UNITS OF THE ARMY OF GOD.

YOU MAY CONFRIM THE FOLLOWING WITH THE F.B.I. THE SANDY SPRING'S DEVICE'S – GELATIN – DYNAMITE – POWER SOURCE 6 VOLT D BATTERY BOXES, DURACELL BRAND, CLOCK TIMER'S. THE MIDTOWN DEVICE'S ARE SIMILAR EXCEPT NO AMMO CAN'S TUPPERWARE CONTAINERS INSTEAD – POWER SOURCE SINGLE 6 VOLT LANTERN

BATTERIES. DIFFERENT SHRAPNEL, REGULAR NAIL'S INSTEAD OF CUTT
NAILS.

THE ABORTION CLINIC WAS THE TARGET OF THE FIRST DEVICE. THE
MURDER OF 3.5 MILLION CHILDREN EVERY WILL NOT BE 'TOLERATED.'
THOSE WHO PARTICIPATE IN ANYWAY IN THE MURDER OF CHILDREN MAY
BE TARGETED FOR ATTACK. THE ATTACK THEREFORE SERVES AS A
WARNING: ANYONE IN OR AROUND FACILITIES THAT MURDER CHILDREN
MAY BECOME VICTIMS OF RETRIBUTION. THE NEXT FACILITY TARGETED
MAY NOT BE EMPTY.

THE SECOND DEVICE WAS AIMED AT AGENTS OF THE FEDERAL
GOVERNMENT I.E. A.T.F., F.B.I., MARSHALL'S E.T.C. WE DECLARE AND WILL
WAGE TOTAL WAR ON THE UNGODLY COMMUNIST REGIME IN NEW YORK
AND YOUR LEGASLATIVE BUREAUCRATIC LACKEY'S IN WASHINGTON. IT IS
YOU WHO ARE RESPONSIBLE AND PRESIDE OVER THE MUR OF CHILDREN
AND ISSUE THE POLICY OF PEVERSION THAT DESTROYING OUR PEOPLE. WE
WILL TARGET ALL FACILITIES AND PERSONNEL OF THE FEDERAL
GOVERNMENT. THE ATTACK IN MIDTOWN WAS AIMED AT THE SODOMITE
BAR (THE OTHERSIDE). WE WILL TARGET SODOMITES, THERE
ORGANIZATIONS, AND ALL THOSE WHO PUSH THEIR AGENDA.

IN THE FUTURE WHEN AN ATTACK IS MADE AGAINST TARGETS WHERE
INNOCENT PEOPLE MAY BECOME THE PRIMARY CAUSALTIES. A WARNING
PHONE CALL WILL BE PLACED TO ONE OF THE NEWS BUREAU'S OR 911.

(Gales, 2010:282)

L1 Text 1

4.2.1.2 The Lampley Hollow letter

No background information is available on this letter.

Hello asshole. This is the eve of the bloodiest day in the history of Lampley Hollow!

You fucks want to step outside the law to show us much of a fuck your mother is? Well,
you have attacked innocent people, and now innocent people will pay, on your behalf. And
a few cops trying to stop us.

Sunday is the final day of Founders Day. On that day a minimum of 20 people will die there.

Here is how it will happen: Your department will receive a phone call ten minutes to the top of an hour, to announce the countdown. At the hour, the first explosion will occur. Approximately six will die, mainly family members, and the bomber. This will start a panic, with people running in all directions. One of those directions will be toward the second bomber. Six seconds after the first explosion the second will occur, a distance from the first. Six more dead.

NOW for the big one. Two groups of people will collide, while escaping their respective explosions. At that time and place the third, largest explosion will occur. Eight dead, at least.

You wonder why we have people willing to do this and die over you? It's because they don't even know they are packing. And you cannot find them.

The people that die will even the score, and we start fresh. Don't fuckup or it will happen again. Perform your job with respect and dignity for the people you serve and you will save their lives. We regret this but feel an example of death is the only way to make you understand.

You remember the bomb in the planter last summer? That's right, the iron pipe bomb, with an electronic igniter. It was powered by four AA batteries in an Electronic supply pack, with a time delay. Don't count on a misfire this time. We worked out the ignition problems with that design.

It's a great day coming.

(Gales, 2010:277)

L1 Text 2

4.2.1.3 Luke Jon Helder, the pipe bomber

Lucas John 'Luke' Helder, also known as the Midwest Pipe Bomber, is an American domestic terrorist and former University of Wisconsin–Stout student from Pine Island, Minnesota. In 2002, while attending the University of Wisconsin–Stout, Helder planned to plant pipe bombs in mailboxes across the United States to create a smiley face shape on the United States map. The bombs, which were packed with BBs and nails, were rigged to explode as the mailboxes were opened. Rigged bombs were found in Nebraska, Colorado, Texas, Illinois and Iowa. In Iowa, six people, including four mail carriers, were injured when the bombs detonated. Ultimately, Helder planted 18 bombs and covered 5,100 km. Notes attached to the bombs went on to denounce government control over daily lives, deny that anyone who had died was really

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

dead, and promised more of the same kind of message. The following is an example of these notes.

Mailboxes are exploding! Why, you ask?

Attention people.

You do things because you can and want (desire) to

If the government controls what you want to do, they control what you can do.

If you are under the impression that death exists, and you fear it, you do anything to avoid

it. (This is the same way pain operates. Naturally we strive to avoid negative emotion/pain.)

You allow yourself to fear death!

World authorities allowed, and still allow you to fear death!

In avoiding death you are forced to conform, if you fail to conform, you suffer mentally and physically. (Are world powers utilizing the natural survival instinct in a way that allows them to capitalize on the people?)

To 'live' (avoid death) in this society you are forced to conform/slave away,

I'm here to help you realize/understand that you will live no matter what!

It's up to you people to open your hearts and minds. There is no such thing as death. The people I've dismissed from this reality are not at all dead.

Conforming to the boundaries, and restrictions imposed by the government only reduces the substance in your lives. When 1% of the nation controls 99% of the nations total wealth, is it a wonder why there are control problems?

The United States strives to provide freedom for their people. Do we really have personal freedom? I've lived here for many years, and I see much limitation. Does the definition of freedom include limitation? I've learned about the history of various civilizations in history, and I see more and more limitation. Do you people enjoy this trend of limitation? If not, change it!

As long as you are uninformed about death you will continue to say 'how high', when the government tells you to 'jump'. As long as the government is uninformed about death they will continue tell you to 'jump' Is the government uninformed about death or are they pretending?

You have been missing how things are, for very long. I'm obtaining your attention in the only way I can. More info is on its way. More 'attention getters' are on the way. If I could, I would change only one person, unfortunately the resources are not accessible. It seems killing a single famous person would get the same media attention as killing numerous un-famous humans. There is less risk of being abducted. Associated with dismissing certain people.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

Sincerely,
Someone Who Cares
PS. More info will be delivered to various locations around the country.

(Olsson, 2008:198-199)

L1 Text 3

4.2.1.4 Theodore Kaczynski (the Unabomber) letter

This case is explained in Section 1.1. The following is the letter sent to the FBI.

This is a message from the terrorist group FC.

We blew up Thomas Mosser last December because he was a Burston-Marsteller executive. Among other misdeeds, Burston-Marsteller helped Exxon clean up its public image after the Exxon Valdez incident. But we attacked Burston-Marsteller less for its specific misdeeds than on general principles.

Burston-Marsteller is about the biggest organization in the public relations field. This means that its business is the development of techniques for manipulating people's attitudes. It was for this more than for its actions in specific cases that we sent a bomb to an executive of this company.

Some news reports have made the misleading statement that we have been attacking universities or scholars. We have nothing against universities or scholars as such. All the university people whom we have attacked have been specialists in technical fields. (We consider certain areas of applied psychology, such as behavior modification, to be technical fields.)

We would not want anyone to think that we have any desire to hurt professors who study archaeology, history, literature or harmless stuff like that. The people we are out to get are the scientists and engineers, especially in critical fields like computers and genetics. As for

156
the bomb planted in the Business School at the U. of Utah, that was a botched operation. We won't say how or why it was botched because we don't want to give the FBI any clues. No one was hurt by that bomb.

In our previous letter to you we called ourselves anarchists. Since 'anarchist' is a vague word that has been applied to a variety of attitudes, further explanation is needed. We call ourselves anarchists because we would like, ideally, to break down all society into very small, completely autonomous units. Regrettably, we don't see any clear road to this goal, so we leave it to the indefinite future.

Our more immediate goal, which we think may be attainable at some time during the next several decades, is the destruction of the worldwide industrial system. Through our

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

bombings we hope to promote social instability in industrial society, propagate anti-industrial ideas and give encouragement to those who hate the industrial system.

The FBI has tried to portray these bombings as the work of an isolated nut. We won't waste our time arguing about whether we are nuts, but we certainly are not isolated. For security reasons we won't reveal the number of members of our group, but anyone who will read the anarchist and radical environmentalist journals will see that opposition to the industrial-technological system is widespread and growing.

Why do we announce our goals only now, through we made our first bomb some seventeen years ago? Our early bombs were too ineffectual to attract much public attention or give encouragement to those who hate the system. We found by experience that gunpowder bombs, if small enough to be carried inconspicuously, were too feeble to do much damage, so we took a couple of years off to do some experimenting. We learned how to make pipe bombs that were powerful enough, and we used these in a couple of successful bombings as well as in some unsuccessful ones.

[Passage deleted at the request of the FBI.]

Since we no longer have to confine the explosive in a pipe, we are now free of limitations on the size and shape of our bombs. We are pretty sure we know how to increase the power of our explosives and reduce the number of batteries needed to set them off. And, as we've just indicated, we think we now have more effective fragmentation material. So we expect to be able to pack deadly bombs into ever smaller, lighter and more harmless looking packages.

On the other hand, we believe we will be able to make bombs much bigger than any we've made before. With a briefcase-full or a suitcase-full of explosives we should be able to blow out the walls of substantial buildings.

Clearly we are in a position to do a great deal of damage. And it doesn't appear that the FBI is going to catch us any time soon. The FBI is a joke.

The people who are pushing all this growth and progress garbage deserve to be severely punished. But our goal is less to punish them than to propagate ideas. Anyhow we are getting tired of making bombs. It's no fun having to spend all your evenings and weekends preparing dangerous mixtures, filing trigger mechanisms out of scraps of metal or searching the sierras for a place isolated enough to test a bomb. So we offer a bargain.

We have a long article, between 29,000 and 37,000 words, that we want to have published. If you can get it published according to our requirements we will permanently desist from terrorist activities. It must be published in the New York Times, Time or Newsweek, or in some other widely read, nationally distributed periodical.

Because of its length we suppose it will have to be serialised. Alternatively, it can be published as a small book, but the book must be well publicised and made available at a moderate price in bookstores nationwide and in at least some places abroad. Whoever

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

agrees to publish the material will have exclusive rights to reproduce it for a period of six months and will be welcome to any profits they may make from it.

After six months from the first appearance of the article or book it must become public property, so that anyone can reproduce or publish it. (If material is serialised, first instalment becomes public property six months after appearance of first instalment, second instalment, etc.)

We must have the right to publish in the New York Times, Time or Newsweek, each year for three years after the appearance of our article or book, three thousand words expanding or clarifying our material or rebutting criticisms of it.

The article will not explicitly advocate violence. There will be an unavoidable implication that we favor violence to the extent that it may be necessary, since we advocate eliminating industrial society and we ourselves have been using violence to that end. But the article will not advocate violence explicitly, nor will it propose the overthrow of the United States Government, nor will it contain obscenity or anything else that you would be likely to regard as unacceptable for publication.

How do you know that we will keep our promise to desist from terrorism if our conditions are met? It will be to our advantage to keep our promise. We want to win acceptance for certain ideas. If we break our promise people will lose respect for us and so will be less likely to accept the ideas.

Our offer to desist from terrorism is subject to three qualifications.

- First: Our promise to desist will not take effect until all parts of our article or book have appeared in print.
- Second: If the authorities should succeed in tracking us down and an attempt is made to arrest any of us, or even to question us in connection with the bombings, we reserve the right to use violence.
- Third: We distinguish between terrorism and sabotage. By terrorism we mean actions motivated by a desire to influence the development of a society and intended to cause injury or death to human beings. By sabotage we mean similarly motivated actions intended to destroy property without injuring human beings. The promise we offer is to desist from terrorism. We reserve the right to engage in sabotage.

It may be just as well that failure of our early bombs discouraged us from making any public statements at that time. We were very young then and our thinking was crude. Over the years we have given as much attention to the development of our ideas as to the development of bombs, and we now have something serious to say. And we feel that just now the time is ripe for the presentation of anti-industrial ideas.

Please see to it that the answer to our offer is well publicised in the media so that we won't miss it. Be sure to tell us where and how our material will be published and how long it will take to appear in print once we have sent in the manuscript. If the answer is

satisfactory, we will finish typing the manuscript and send it to you. If the answer is unsatisfactory, we will start building our next bomb.

We encourage you to print this letter.

FC

[Passage deleted at the request of the FBI.]

(Farhi, 2015; Kaczynski, 1995)

L1 Text 4

4.2.2 Second language English speakers' texts

Professor Hilton Hubbard testified in a South African extortion case in 1989. (Hubbard, 1994; 1995). This case was based on a review of ten extortion letters sent to the Johannesburg branch of a national supermarket chain. The letters threatened to poison food on the company's shelves and then inform the press unless R1.5 million (about R6 million now) was paid. Because the defendant was a native Polish speaker with a less-than-perfect command of English, an error analysis was conducted, by Hubbard, on two sets of texts: (i) the letters received by the company and (ii) a total of seven essays of varying lengths written by English second-language speakers, one of whom was the defendant. While it was obvious that an error analysis alone would not be enough to get a conviction (SCSA, 1989), Hubbard emphasised that the error patterns in the accused's writings must be quantitatively compared with those of a person with a comparable background and level of English proficiency.

Six of the seven essays, as mentioned above, were made available for this study's analysis by Professor Hubbard. The actual extortion letters were not accessible, which made it impossible to compare the essays with the letters. Four of the essays were selected by the researcher for this study, so that there was an equal number of L1 and L2 texts to compare. The four texts are presented below.

4.2.2.1 'My Work'

Why do I work?

- I use acquired knowledge

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

- I have also a degree of satisfaction from my work achievement.
- necessity – as the salary which I earn allows me to live, travel and provides for my future when I stop to work.

Usually a working person spends three quarters of his useful life at work. The person is involved with work problems for most of his useful, productive life. After the person's education is completed or a certain level of knowledge is reached, which allows to start work, till the retiring age, all of life is dominated by work and its problems. For the majority of people this period last for thirty to fourty years.

During this time a person also builds his family, his life, outside interests: his whole reason for living, of course these can only be achieved if the work situation is normal, the income from work is steady and sufficient for all the expenses required to continue with one's life on a desired level.

In my opinion it is extremely important that the performed work gives the person satisfaction and enough pleasure to continue with it, without too much stress.

On the other hand the performed work must be demanding enough to force the person to develop further.

My work is mostly done in the same field, electricity, with a number of changes during the years.

I was and I am involved in electrical engineering heavy current and high voltage.

However, I was involved with some aspect of light current, as for a number of years I was involved in the electrical protection. Design for the protection for the substation rural or urban, testing the designed protection scheme and commissioning.

My work gives me a degree of satisfaction but also some frustration.

I feel that there is a number of reasons for this:

- I am a woman who works in a man's world,
- a certain degree of defficiency in my working style is not allowing me to perform the work excellently,
- the poor comandments of English languages, which is not my mother-tongue and Im not educated in it.

It is enough for an introduction. It is the time to say a few words about my work as such. After receiving my degree as an electrical engineer I completed my 'apprentice' training as an 'engineer in training' which is a practical training, in the high voltage laboratory of the transformer factory in Poland.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

It was my first job and really highly challenging. The transformers produced there were of a new design and they required a full type test to be accepted by the client for required purposes.

The laboratory was also brand new, with special equipment from which the most important was an impulse generator, capable of producing 2,1 million volts for a number of microseconds. [sic]

The team of people working there consisted of a number of young engineers, mostly friends, as we completed our studies in the same technical university.

Everybody there was very enthusiastic and dedicated and also our leader, our professor, knew how to deal with the young enthusiastic crowd of workers.

The situation became different when I started work in South Africa, after receiving some extra training in England at English Electric in Stafford.

When I look back, I can say honestly, I did not do too bad, as I was accepted by the fellow engineers, maybe as a kind of 'oddy', and also by the bosses I worked for.

My present work covers quite a broad field of the generator and auxiliaries and the generator system.

Due to the different manufactures of the electrical generators which were installed over a number of years, in various power stations, I deal with quite a spectrum of different designed machines.

Generally the generators differ in their basic size, generated power, cooling system, insulation material and in the regulation system. The principle of generating the electrical power of course is the same for small and big generators.

The generators installed in the older power stations have they MCR – (maximum continuous rating) of 30 MVA and in the new one 680 MVA and in the nuclear power station 920 MVA.

The cooling systems of course are different; the cooling median used are: air, hydrogen and water.

These require that the construction of the stator bars, which carry generating current and situated in the stator iron slot are different. As the cooling system must be more efficient for the higher current produced by the generator.

Each manufacturer tries to introduce his own specific, patented, design.

The aspects which make my work interesting are:

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

- a) setting a specification for the generator which includes requirements for the future stations and equipment installed in them.
- b) checking an answer, in a form of full proposal, to a tender issued and recommending, from the technical point of view, the most successful tenderer, and specifying why. However the financial aspect of the tender is also taken into consideration.

L2 Text 1

4.2.2.2 'JOB RELATED OR WORK RELATING STORY'

DURING MY EMPLOYMENT WITH A CONTRACTING FIRM CALLED 'ENGINEERING DRAFTING SERVICES' I WAS APPROACHED WITH MY AGENCY, IF I WOULD NOT LIKE TO CONSIDER AND FOR ESKOM, BECAUSE THEY HAVE BEEN ADVERTISING A VACANT POST IN THE LIGHTING SECTION.

THOSE TIME I HAVE HAD A SHORT TIME-TERM CONTRACT WITH A CONSULTING ENGINEERING WHICH WAS LEADING TO THE END. WHILE I WAS EMPLOYED WITH THE CONSULTING ENGINEERS I HAVE BEEN INVOLVED ALSO IN THE LIGHTING FIELD FOR DOMESTIC INSTALLATION SUCH AS: SHOPPING CENTRE, DUPLEXES, SCHOOLS, HOTELS, PRIVATE RESIDENCES. WHEN I HEARD THAT I COULD HAVE A LONG TERM CONTRACT WITH SUCH A BIG COMPANY AS ESKOM I WAS A LITTLE FRIGHTENED, BUT I HAD ACCEPTED ALSO BECAUSE I KNEW FEW CZECHOSLOVAKIAN PEOPLE THERE. IT ALL HAPPENED EXACTLY IN FEBRUARY 1978, WHEN I STARTED WORKING FOR ESKOM IN MEGAWATT PARK, SITUATED IN SANDTON-RIVONIA. THE COMPLEX WAS SO BEAUTIFUL IT HAD FASCINATED EVERY NEWCOMER. THE WORKSTATION WAS ALL OPEN PLAN OFFICES WITH VERY MODERN FURNITURES AND MOST BEAUTIFUL PLANTS I HAVE EVER SEEN IN MY LIFE. THE SECTION WHICH I HAVE JOIN WAS FAIRLY BIG AND IT WAS RESPONSIBLE FOR THE DESIGN OF THE WHOLE POWER STATION LIGHTING, WHICH CONSISTED OF 1) INTERIOR LIGHTING OF BUILDINGS SUCH AS OFFICES, CONTROL ROOMS, LECTURE ROOMS, DINNING AREAS, LABORATORIES, RECEPTION AREAS, WORKSHOPS AND WILD VARIETY OF OTHER BUILDING LIGHTING. WITH THIS KIND OF WORK I WAS QUITE FAMILIAR AND I FELT CONFIDENT AND

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

COMPLETED MY WORK WELL TO MY SUPERVISOR SATISFACTION. BUT WHEN IT COME TO DO THE LIGHTING DESIGN OF THE OUTSIDE PLANTS SUCH AS TURBINE HOUSE, BOILERS HOUSE, PRECIPITATORS, COAL AND ASH DRIVE HOUSES AND TRANSFER HOUSES AND THEIR RELEVANT CONVEYORS RUNS. THOSE ASSIGNMENTS WERE VERY MUCH MORE COMPLICATED AND REQUIRED ALOT OF ENGINEERING INPUT. THE LIGHTING DESIGN OF THOSE PLANTS INVOLVED REGULAR SITE VISITS, CLEAR AND UNDERSTANDABLE COMMUNICATION WITH ALL RELEVANT DISCIPLINES INVOLVED IN THE PROJECT SUCH AS ARCHITECTURAL, MECHANICAL, CIVIL A/C-ING, TELECOMUNICATION AND OF COURSE THE CONTRACTORS DISCIPLINES. IT ALSO REQUIRED THE STUDY OF HOW THE PLANTS FUNCTIONING, STUDY OF THE CONSTRUCTION DRAWINGS, ABILITY TO READ THE MAKERS DRAWING TO BE ABLE TO PRIPARE A GENERAL ARRANGEMENT DRAWINGS OF THE PLANTS AND THEN THE REQUIRED LIGHTING FOR GOOD VISIBILITY, GOOD SECURITY AND SAFETY AND ALSO BEUATY AND EMPHASIS ON THE PLANT. THE PROPER LIGHTING DESIGN COULD ONLY BE ACHIEVED WITH A SYSTEMATIC APPROACH, PROPER SITE MEASUREMENTS AND SITE MARK-UPS AND OBTAINING THE ALL UP-TO DATE INFORMATIONS. THIS TYPE OF WORK WAS VERY INTERESTING, CHALENGING AND VERY PROMISING. THE TYPE OF WORK, WHICH I WAS INVOLVED IN, AND THE LOVELY PEOPLE I WORKED WITH, AND A PLEASANT ENVIRONMENT AND ATMOSPHERE MADE ME TO CHANGE MY MIND AND BECOME A PERMANENT EMPLOYEE TO ESKOM. THEN IN OCTOBER 1978 I STARTED ON THE PERMANENT BASIS OF A SENIOR DRAUGHTSWOMAN AND ALL THE PEOPLE EXCEPTED ME WITH WARM WELCOME TO BECOME ONE OF THEM. AS THE TIME PASSED BY, I HAVE GAINED MORE EXPERIENCE AND ALSO GOT INVOLVED IN MORE DIFFICULT ASSIGNMENTS. AFTER MANY OF SUCCESFUL JOB COMPLETIONS ON THE COAL POWER STATION I RECEIVED A COMPLIMENTS WHICH MADE ME EVEN MORE ENTHUSIASTIC AND MORE INTERESTED IN MY WORK. AFTER FEW YEARS OF WORKING ON THE COAL POWER STATION I HAVE BEEN ASKED FROM MY DIVISION MANAGER IF I WOULD NOT TO LIKE TO GET INVOLVED IN THE LIGHTING DESIGN FOR A

HYDRO STATION. I DID EXCEPT THE NEW CHALLENGE AND STARTED WORKING ON THE DRAKENSBERG PUMPED STORAGE SCHEME. I MUST SAY IT WAS QUITE AN ADVENTURE AND FAN IS WELL, WHEN WE VISITED THE SITE WE HAD TO CHANGE TO THE RUBBER CLOTHING RUBBER BOOTS BECAUSE THE TUNNELS THROUGH WHICH WE HAD TO ACCESS THE UNDERGROUND ENVIRONMENT WAS ALWAYS FULL OF WATER, DURING THE EXCAVATION. WE REALY LOOKED LIKE PEOPLE FROM THE MOTHER PLANET, MEANWHILE WE WERE WORKING UNDER THE GROUND COUPLE OF HUNDRED METERS. I WAS INVOLVED IN THIS HYDRO PROJECT RIGHT TO THE END AND ALSO IT BECAME ONE OF MY BIGGEST SUCCESSFUL ACHIEVEMENTS. AFTER THE COMPLETION OF THIS PROJECT I HAVE BEEN PROMOTED TO THE DESIGN DRAUGHTSWOMAN OFFICIALLY, EVEN THE RESPONSIBLE DUTY I HAVE BEEN CARRYING OUT LONG TIME ALREADY. IT WAS A VERY NICE AND PLEASANT FEELING AND I WAS VERY PROUD OF IT. THE WORK SITUATION HAD CHANGED A BID BUT MAJORITY WAS STILL THE SAME. I WAS MORE INVOLVED FOR PREPARATION WORK FOR THE WHOLE SECTION, MORE DESIGN ORIENTATED FOR THE INDIVIDUALS AND ALSO WAS INVOLVED IN ESKOM GENERAL DATA FOR STANDARDS. MY WORK BECAME MORE ADMINISTRATIVE. THEN IN 1986 I HAVE ATTENDED AN ILLUMINATION ENGINEERING COURSE HELD AT TECHNIKON AND ORGANISED BY 'ILESA', WHICH I COMPLETED SUCCESFULLY. I MUST ADMIT IT, THAT ONLY AFTER I COMPLETED THIS COURSE I HAVE FOUND OUT HOW MUCH MORE THE WORD OF LIGHTING REALY MEANT, HOW MANY SECRETS LIE IN THE DESIGN AND HOW LITTLE I REALY KNEW. I MUST SAY THAT IT CHANGED MY OWN APPROACH TO THE DESIGN PARAMETERS FOR THE LIGHTING. DURING THE YEAR OF THE COURSE WE LEARNED THE PHYSIOS OF THE HUMAN EYE, THE SPECTRUM, PHOTOMETRY, COLOUR REENDERING, LAMP AND LUMEN STRUCTURES, STREET LIGHTING, ROAD LIGHTING, CHURCHES, MUSLUM AND GALLERIES, SPORT FLOODLIGHTJG DAYLIGHTING, SUPPLEMENT ARTIFICIAL LIGHTING AND MANY MANY MORE INTERESTING AND FASCINATING FACTORS NEEDED TO A PROPER LIGHTING DESIGN. AFTER I COMPLETED THIS COURSE I HAVE TRIED TO

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

TRANSFER OR MY NEW EXPERIENCE ON TO MY COLLEGES. IT WAS DIFFICULT IN THE BEGINING BUT SLOWLY BECOME RECOGNISIBLE. THEN A YEAR LATER IN MARCH 1987 I WAS PROMOTED TO THE SECTION LEADER OF THE LIGHTING SECTION, WHICH I HELD UP TO NOW AND MY JOB MISSION IS: TO PROVIDE TECHNICAL SERVICES IN THE FIELD OF THE LIGHTING AND SMALL POWER FOR VARIOUS POWER STATIONS: MAJUBA, KENDAL, MATIMBA

L2 Text 2

4.2.2.3 No Title

Gentlemen

I must admit that I read your newspaper rather occasionally and when I do so I am rather disapointed. All the articles on the first page discribe sensational events which are rather of a mean or no value to me and to other readers as well, I hope. Is it done on purpose? Is this a vay to atract attention to somthing less important and make the readers forget about some real problems in our country?

Have you been on Church street during lunch time recently? Have you seen all the baggers? What have you done to help them to change their lives?

Should we not talk about special schools for black disable people? Even they can be useful in a society but we must create an opportunity for them to lern them a trade. It does not to be something complicated but simple and useful. It will improve their quality of live and ours at the same time.

With regards

On the day of my retirement I will pack my suitcases and go to my country side house.

My country side house will be an old farm house with a garden full of flowers. All the goods inside my house will be made of a wood and other natural materials. There will not be place for any plastic!

During my retirement I would like to do all the things for which I have not had time befor. Books reading, pottery, drawing perhaps even painting. But, I would like to have somebody to talk with, to exchange opinions and to argue. To do that will be a great pleasure but I would like to be somhow useful to other people as well. How? I do not know yet but I have still some time to think about it.

Pollution – this is one of the biggest problems of our century and one should think what price we are going to pay for it.

But even bigger problem is that so many people are not aware of the consequences of the pollution. Many others are aware but money are more important for them. Unfortunately in SA does not exist any law which will force the industry to take some precaution measures.

4. Conservation

Every generation should think about other generations which will come after and which will have to live in the world the previous generation have left. We should use the nature to our advantage but not damage it. We should let our children and grandchildren enjoy it as well. Use does not mean abuse therefore everything in the nature should be used sparingly and a great care should be taken to restore that what was used. Now days we are able to predict the consequences of our activities and more attention should be paid to plan a proper method to diminish the negative sides of existence. It sounds very pompous!

I am afraid that I can not think of anything that will keep the prices of food down. Should the government subsidise the food? The high prices of food production are connected to the high prices of anything else e.g. transport costs, labour etc. Are the big supermarkets a good solution for a food distribution? Will small shops not be better? The running cost of a supermarket must be very high and a cost of wasted (too old for sale) food must be included in the price. Small shops are normally more elastic in a decision making process, better organised and less food is wasted in them. A good developed network of small shops should create more competition, better quality and lower prices. The government should do something to make people to lead their lives on a level they can afford. The very high inflation is between others due to big credits people live on, due to spending money which do not exist. But this does not concern only people in the country but the country itself. What did happen to SA economy? Do we export enough goods to import everything we need?

Another problem is the society awareness and understanding of economical problems. This is a big role for TV, radio and press to play.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

I start every day at 5h45 with physical exercises. I try to make myself tired and it takes me about 15 minutes. Then, according to my son, I run around doing nothing a half asleep a half awaked. But still during this time I prepare a breakfast for my son and me, dress myself and at 7h00 I am ready to leave for my work. I start work at 7h45 with checking my diary where I write all appointments and things to be done on that day. I make a lot of effort to work until lunch time. But not every day I have time for lunch time and then I work through until afternoon teetime.

My work ends normally at 16h45. After work I rush home, then prepare food for my son who is always very hungry. At home I start the second shift of my work. I play tennis on Mondays, Thursdays and sometimes Fridays.

L2 Text 3

4.2.2.4 *No Title*

Dear Sirs,

It remains an undeniable fact that there are millions of hungry people in our country. It is the result of commonly spread inconsistency of the white race: to help to natives (by means of food supply, medical care), to make them grow in number (i.e. to destroy the natural rate of population growth), and then to try to help the hungry artificially multiplied population. Of course we are far from admittance to the guilt: oblivious to the real reason which produces millions of the poor and hungry we try to do our hypocritical best to help them by throwing odds and ends from our fully packed fridges. But still as long as we do that out of guilty conscience we are being punished enough. The only solution to the problem which comes to my mind is as surrealistic as the cause of the problem: why not to supply the hungry with invisible caps so they could come sometimes to our homes, eat to their heart contents without arising our righteous indignancy to such a outrageous behaviour.

The day when I retire appears to be to most glamorous, the most dreamt of day of my life. I think I have so many expectations, planned activities, a general feeling of long awaited freedom to do whatever I please that it all makes difficult to define something in detail. Which is wise since I shall probably change and so will my interests. Had the day be today I would not have been writing this paper for Hilton (for sure) but I would be bursting with joy tossing in my garden, fondling flowers, harrng up the buds, trying to

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

communicate with grass and trees and clouds, guessing their wishes and needs and trying to satisfy them. But of course things may happen: I may fall ill (what about quite common at retiring age Alzheimer disease?) or in any other way disabled and then it is better to be right where I am today.

In general I think that pollution is the thing which should not have happened at all (which means that perhaps the three last centuries should not have happened).

I have been impressed by learning the theory about GAJA (the great goddess of Earth) which assumes that earth is a living organism. Although the idea seems to be shocking (to say the least) the scientists who favor it were able to find so many indisputable proves to support it (the same constant temperature around the earth from billions of years for instance) that no reasonable human being does not reject it any longer.

When I think about pollution I have to ask the following question on behalf of Gaje: is there a safe place from mankind?

Conservation should be treated as the most vital of our activities if we wish to survive as a species. We should become more humble towards nature. We should restrain at once our needs which seem to be blown up beyond every reasonable limit (meeting our hyper needs we tend to produce more at the expense of more and more robbed and violated nature). We should undertake any industrial (i.e. potentially disastrous to nature) activity only when we are able to neutralize the harmful to our environment effects. Then and only then we might be able to think that the future generations will have fresh air to breathe, nutritious healthy food to eat and clear water to drink (to say nothing about marvellous natural surroundings to live in).

The first solution which immediately crosses one's mind is: to take from the rich and to give to the poor; but this would be a very shortlasting remedy. (And of course in the end there would be the newly rich and the newly poor so the whole process should have to be repeated ceaselessly, which seems to be boring and tiresome. What perhaps could be a solution is the existence of a honest, dedicated and wise government with a sound economic policy. Do you know of one?

The high prices would eventually affect the whole population although it seems obvious that those who are interested in buying a Sunflowers by Van Gogh at R50 000 000 would suffer slightly less than the rest of society.

Here it is : let's have an agreement with the price levels (or let's force them) not to change the digits they bear with time. Let's make the labels indifferent to any, even the most considerate or on the contrary: most cruel, measures undertaken in order to change what they have read from their birth. And they could even learn how to cheat the oppressors: at first they would pretend to have the new numbers printed on them but the moment the price putting men goes away the labels would assume the previous price. Our future prosperity lies in the ingenuity of price levels!

My weekly schedule seems to be unbearably boring. The days are so enjoyably similar to each other. But if you see what are they filled up with you will understand why I have said 'enjoyably similar'. Every working day I get up at six and immediately check on the weather which is silly because I know I will enjoy any kind of the sky mood. Then if it not raining I go out and run for a couple of minutes in the garden greeting everything what I can see on my way. Then I get ready to go to work which is very creative since I have nothing to put on but still I have to dress.

Every day I take a different way to work; I mean physically it's the same streets but not by what I see. Every day and every moment of my driving ups and downs I see different hills and woods, roofs, windows and magnolias, different horizons and absolutely amazingly different colours of everything. Then I work with greater or lesser pleasure which depends on my mood and also on the subject I am busy with.

L2 Text 4

4.3 Feature extraction

Below is a description of how each characteristic was identified, tagged and quantified. As indicated in Section 2.8, the writing-style characteristics are: (i) average sentence length; (ii) internal structure of sentences; (iii) end-of-sentence punctuation; (iv) T-units (average length and quantity per sentence); and (v) cohesion markers (reference, substitution, ellipsis, conjunction and lexical cohesion). The texts and the manual analyses are attached as Appendices One to Eight for reference.

4.3.1 Average sentence length

Using Microsoft Word, each text was highlighted to reveal its total number of words. The number of sentences was carefully tallied. The average sentence length was calculated by dividing the total number of words by the number of sentences, as the following formula shows.

Total amount of words in text ÷ total amount of sentences in text = Average sentence length.

$$\frac{\text{Total amount of words in text}}{\text{Total amount of sentences in text}}$$

4.3.2 Internal structure of sentences

Each sentence's clauses were marked by hand with a purple pen, with each clause enclosed in brackets. The order of the words was also marked to answer the question: Does the writer follow the typical syntactical order in the English language, of SVN (subject – verb – noun)?

4.3.3 End-of-sentence punctuation

The writers' use of punctuation at the end of sentences was marked by hand with a red pen and noted in writing.

4.3.4 Average T-unit length

The main T-unit in each sentence was highlighted by hand with a yellow marker. The number of words within a T-unit was then counted and added to receive the total number of words in a T-unit. This was then divided by the number of sentences. As mentioned in Section 2.10.2, the length of a T-unit conveys syntactic complexity, meaning that if the length of a writers' T-units is almost as long as the sentence itself, it is more syntactically complex. In other words, if the T-unit length is as great as the average sentence length (or close to it), the writer is more proficient in the use of the language. The equation is shown as follows: Total amount of words in the T-units ÷ Total number of sentences in the text = Average T-unit length.

$$\frac{\text{Total amount of words in the T – units}}{\text{Total number of sentences in the text}}$$

4.3.5 Average number of T-units

The number of T-units within each sentence of each text was indicated by hand with a green pen.

4.3.6 Reference

As discussed in Section 2.10.3.1, there are four types of reference: particular reference, no particular reference, generic reference and unique reference. A blue pen was used to identify all the instances of reference in each text, and the type of reference was noted.

4.3.7 Substitution

As discussed in Section 2.10.3.2, substitution is the replacement of one element by another element (a substitute word or expression) in such a way that the meaning of the sentence is not affected. Each instance of substitution was marked by hand with a blue marker, and then the number of substitutions was checked with the concordance feature of *WordSmith Tools*.

4.3.8 Ellipses

Ellipsis (plural ellipses; also referred to in the literature as contraction, continuance or deletion) involves the omitting of certain elements from a passage because the omission streamlines the writing and does not affect meaning (Carstens & Van de Poel, 2012) – as discussed in Section 2.10.3.3. The instances of ellipsis were marked by hand with a pink marker, with the symbol ^ used to indicate the omission. Since ellipses are not common in texts, the number was counted manually and written down.

4.3.9 Conjunctions

As also discussed in Section 2.10.3.4, conjunctions are common in texts, enhancing flow and readability and thereby achieving coherence. Conjunction markers link preceding and subsequent sentences or phrases, indicating the nature of the relationship between the two parts (Carstens & Van de Poel, 2012). An orange marker was used to mark each instance of conjunction by hand and the numbers were then tallied with the concordance feature of *WordSmith Tools*.

4.3.10 Lexical cohesion

Lexical cohesion concerns the linkages established between so-called content words (nouns, verbs, adjectives and adverbs) (Carstens & Van de Poel, 2012). Lexical cohesion occurs in two forms: repetition and collocation (Carstens & Van de Poel, 2012). Repetition occurs through the direct repetitions of words. Collocation refers to words associated with each other at the meaning level, even if the nature of the relationship is not entirely clear. Types of lexical repetition are: a physical repetition of the same item; synonyms; superordinates; and epithets. Types of collocation are: synonyms; meaning opposition and contrast; hyponyms, part-whole relationships; ordered and unordered series; and syntagmatic relations. All instances of lexical cohesion were tagged by hand with a green marker, with the type noted in writing.

Table 4.3 presents the features extracted from the texts with the corresponding colours used to highlight each.

Table 4.3: Feature extraction colour coding

<i>Feature</i>	<i>Colour</i>	<i>Identifier</i>
Internal structure of sentences	Purple pen	[]
Main T-unit	Yellow marker	Highlighted
Number of T-units per sentence	Green pen	Written after sentence
Reference	Blue pen	Circled
Substitution	Blue marker	Underlined
Ellipse	Pink marker	^
Conjunction	Orange marker	[]
Lexical cohesion	Green marker	Underlined

4.4 Findings

Table 4.4 presents the results of the analysis of the L1 texts, and Table 4.5 presents the results of the analysis of the L2 texts.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

Table 4.4: Results of L1 texts' analysis

L1 Texts										
	Chaski's Markers			T-units		Cohesion Markers				
	Average Sentence Length	Internal Structure of Sentences	End-of-Sentence Punctuation	Average T-unit Length	Average Amount of T-units	Reference	Substitution	Ellipse	Conjunction	Lexical Cohesion
L1 Text 1	264 ÷ 16 = 16,5 words per sentence	Average of 2 clauses per sentence SVN	Only full stops	215 ÷ 16 = 13,44 words per T-unit	Range: 1 - 3	Particular reference No particular reference Generic reference Unique reference	7	7	12	Physical repetition of the same item Part-whole relationships
L1 Text 2	327 ÷ 31 = 10,55 words per sentence	Average of 2 clauses per sentence SVN	Full stops Exclamation marks Question marks	235 ÷ 31 = 7,60 words per T-unit	1 or 2	Particular reference No particular reference Generic reference Unique reference	6	3	16	Physical repetition of the same item Superordinate and hyponyms Syntagmatic relations Meaning opposition and contrast
L1 Text 3	418 ÷ 37 = 11,3 words per sentence	Between 1 and 2 clauses per sentence SVN	Full stops Exclamation marks Question marks	367 ÷ 37 = 9,92 words per T-unit	1 or 2	Particular reference No particular reference Generic reference Unique reference	1	6	14	Physical repetition of the same item Synonyms Meaning opposition and contrast
L1 Text 4	1414 ÷ 76 = 18,61 words per sentence	Between 1 and 3 clauses per sentence SVN	Full stops Question marks	1046 ÷ 76 = 13,76 words per T-unit	Range: 1 - 3	Particular reference No particular reference Generic reference Unique reference	7	12	54	Physical repetition of the same item Superordinate and hyponyms Synonyms Meaning opposition and contrast Part-whole relationships Syntagmatic relations

As may be seen in Table 4.4, a number of parallels and contrasts are evident among the four texts. The average sentence length ranges from 10 to 19 words; clauses do not exceed three per sentence; all authors adhere to the standard sentence structure of subject – verb – noun; end-

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

of-sentence punctuation varies from author to author; words per T-unit range from 7 to 14; T-units per sentence range from one to three; all authors use all four types of reference; and each author uses cohesion markers differently.

Table 4.5: Results of L2 texts' analysis

L2 Texts										
	Chaski's Markers			T-units		Cohesion Markers				
	Average Sentence Length	Internal Structure of Sentences	End-of-Sentence Punctuation	Average T-unit Length	Average Amount of T-units	Reference	Substitution	Ellipse	Conjunction	Lexical Cohesion
L2 Text 1	831 ÷ 39 = 21,31 words per sentence	Clauses range from 1 to 6 SVN	Full stops	402 ÷ 39 = 10,31 words per T-unit	Range: 1 - 3 Instances of 4, 6 and 7	Particular reference Generic reference Unique reference	9	4	39	Physical repetition of the same item Superordinate and hyponyms Meaning opposition and contrast Part-whole relationships Syntagmatic relations
L2 Text 2	950 ÷ 36 = 26,39 words per sentence	Clauses range from 2 to 5 Changes order of SVN	Full stops	337 ÷ 36 = 9,36 words per T-unit	Range: 2 - 4	Particular reference Unique reference No particular reference	8	15	54	Physical repetition of the same item Synonyms Part-whole relationships Syntagmatic relations
L2 Text 3	855 ÷ 57 = 15 words per sentence	Clauses range from 1 to 3 SVN	Full stops Exclamation marks Question marks	500 ÷ 57 = 8,77 words per T-unit	Mostly 2	Particular reference Generic reference Unique reference	5	10	38	Physical repetition of the same item Superordinate and hyponyms Synonyms Meaning opposition and contrast Part-whole relationships Syntagmatic relations Ordered series
L2 Text 4	1003 ÷ 37 = 27,11 words per sentence	Clauses range from 2 to 5 Changes order of SVN	Full stops Question marks	458 ÷ 37 = 12,38 words per T-unit	Range: 2 - 4	Particular reference No particular reference Generic reference	5	12	41	Physical repetition of the same item Superordinate and hyponyms Synonyms Meaning opposition and contrast

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
 ATTRIBUTION IN L1/2 ENGLISH TEXTS

						Unique reference				Part-whole relationships
										Syntagmatic relations
										Ordered series

Table 4.5 indicates several areas of commonality and difference between the four authors. The average sentence length ranges from 15 to 28 words, and each sentence contains anywhere from one to seven clauses. Two authors adhere to the standard construction of English sentences, while the other two reverse the order of their clauses. The use of end-of-sentence punctuation varies from text to text; the number of words per T-unit ranges from 8 to 13; the average number of T-units per sentence ranges from 2 to 4; each text makes use of a different type of reference, but only one used all four; the remainder of the cohesion markers are also varied.

4.5 Feature extraction from texts: Some examples

For examples of end-of-sentence punctuation, internal structures of sentences, number of T-units per sentence and average T-unit length, refer to Appendices One to Eight. There are many more examples of each writing style feature than the following shown in Tables 4.6 and 4.7; a selection has been made for illustration purposes.

Table 4.6: Examples of features extracted from L1 texts

L1 Texts					
Cohesion Markers					
	Reference	Substitution	Ellipse	Conjunction	Lexical Cohesion
L1 Text 1	<p>Particular reference: <i>We</i> refers to the group <i>Army of God</i>; <i>their</i> refers to <i>Sodomites</i></p> <p>No particular reference: <i>ammo</i>; <i>facility</i>; <i>sodomites</i></p> <p>Generic reference: <i>the federal government</i>; <i>devices</i>; <i>targets</i>; <i>shrapnel</i></p> <p>Unique reference: <i>Sandy Springs</i>; <i>the FBI</i>; <i>911</i>; <i>New York</i></p>	<p><i>Facility</i> substitutes <i>The abortion clinic</i></p> <p><i>The murder of children</i> substitutes <i>abortion</i></p>	<p><i>We declare and [we] will wage total war...</i></p>	<p><i>and</i></p> <p><i>except</i></p> <p><i>first</i></p> <p><i>second</i></p> <p><i>In the future</i></p>	<p>Physical repetition of the same item: <i>bombings</i>; <i>the FBI</i>; <i>device</i>; <i>murder</i></p> <p>Part-whole relationships: <i>shrapnel</i>, <i>batteries</i>, <i>timer</i>, <i>ammo</i> as parts of the whole, <i>bomb</i></p>

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

L1 Text 2	<p>Particular reference: <i>they</i> refer to the bombers; <i>we</i> refer to the bombers</p> <p>No particular reference: <i>time; place; asshole; fucks</i></p> <p>Generic reference: <i>that day; directions; that design</i></p> <p>Unique reference: <i>Lampley Hollow; Sunday; AA batteries</i></p>	<p><i>That day</i> substitutes <i>Sunday</i></p> <p><i>The big one</i> substitutes the <i>third, largest explosion</i></p>	<p><i>...with respect and [with] dignity...</i></p>	<p><i>and</i></p> <p><i>final</i></p> <p><i>now</i></p> <p><i>at least</i></p> <p><i>at that time and place</i></p>	<p>Physical repetition of the same item: <i>innocent people; explosion; day</i></p> <p>Superordinate and hyponyms: <i>day x Sunday; family x mother</i></p> <p>Syntagmatic relations: <i>hour x seconds; bomb x explosion</i></p> <p>Meaning opposition and contrast: <i>live x death</i></p>
L1 Text 3	<p>Particular reference: <i>you</i> refer to you people; <i>they</i> refer to the government</p> <p>No particular reference: <i>society; boundaries; restriction</i></p> <p>Generic reference: <i>the government; world powers; the nations; the resources</i></p> <p>Unique reference: <i>The United States</i></p>	<p><i>It</i> substitutes <i>this trend of limitation</i></p>	<p><i>...open your hearts and [your] minds...</i></p> <p><i>World authorities allowed and [world authorities] still allow...</i></p>	<p><i>because</i></p> <p><i>and</i></p> <p><i>in a way</i></p> <p><i>as long as</i></p> <p><i>unfortunately</i></p>	<p>Physical repetition of the same item: <i>the government; control; jump; limitation</i></p> <p>Synonyms: <i>realize x understand; boundaries x restriction; kill x dismiss</i></p> <p>Meaning opposition and contrast: <i>live x death; freedom x limitations; famous x un-famous</i></p>
L1 Text 4	<p>Particular reference: <i>we</i> refer to the terrorist group <i>FC</i>; <i>you</i> refer to the <i>FBI</i></p> <p>No particular reference: <i>that bomb; this company; the article</i></p> <p>Generic reference: <i>universities; publications; the authorities</i></p> <p>Unique reference: <i>Burston-Marsteller; December; the FBI; United States Government</i></p>	<p><i>university people</i> substitute <i>scholars</i></p> <p><i>that end</i> substitutes <i>violence to the extent that it may be necessary</i></p>	<p><i>...like computers and [like] genetics...</i></p> <p><i>...the size and [the] shape...</i></p>	<p><i>because</i></p> <p><i>and</i></p> <p><i>but</i></p> <p><i>such as</i></p> <p><i>regrettably</i></p> <p><i>on the other hand</i></p> <p><i>alternatively</i></p>	<p>Physical repetition of the same item: <i>industrial system; botched; bombs; sabotage</i></p> <p>Superordinate and hyponyms: <i>media x book, article, news, manuscript; bombs x gunpowder bombs, pipe bombs</i></p> <p>Synonyms: <i>professors x scholars; briefcase x suitcase; blew up x explosives</i></p> <p>Meaning opposition and contrast: <i>successful x unsuccessful; past x future; harmless x deadly</i></p> <p>Part-whole relationships: <i>bomb x powder, batteries, pipe;</i></p> <p>Syntagmatic relations: <i>company x executive, manager; university x scholar, professor, study; critical fields x computers, genetics; study x archeology, literature, history</i></p>

Table 4.7: Examples of features extracted from L2 texts

L2 Texts					
Cohesion Markers					
	Reference	Substitution	Ellipse	Conjunction	Lexical Cohesion
L2 Text 1	<p>Particular reference: <i>his</i> refers to a working person; <i>it</i> refers to <i>English languages</i></p> <p>Generic reference: <i>electricity; work; transformers; generators; cooling systems</i></p> <p>Unique reference: <i>South Africa; England; English Electric; Stafford</i></p>	<p><i>this period</i> substitutes <i>life is dominated by work and its problems</i></p> <p><i>this</i> substitutes <i>My work gives me a degree of satisfaction but also some frustration.</i></p>	<p><i>...the salary which I earn allows me to live, [the salary which I earn allows me to] travel and [the salary which I earn allows me to] provide for my future...</i></p>	<p><i>also</i></p> <p><i>as</i></p> <p><i>and</i></p> <p><i>till</i></p> <p><i>however</i></p>	<p>Physical repetition of the same item: <i>work; engineer; transformer; generator; cooling systems; manufacturer</i></p> <p>Superordinate and hyponyms: <i>feelings x satisfaction; pleasure; stress; frustration; enthusiastic</i></p> <p>Meaning opposition and contrast: <i>satisfaction x frustration; small x big</i></p> <p>Part-whole relationships: <i>cooling systems x air, hydrogen, water</i></p>

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

					Syntagmatic relations: <i>electricity x volts, voltage, power, current; mother-tongue x English language</i>
L2 Text 2	<p>Particular reference: <i>they</i> refer to <i>Eskom</i>; <i>it</i> refers to <i>the complex</i>; <i>them</i> refers to <i>all of them</i></p> <p>Unique reference: <i>Eskom; Czechoslovakian; Megawatt Park; Sandton-Rivonia; Drakensberg pumped storage scheme</i></p> <p>No particular reference: <i>the lighting section; this project; the contractors</i></p>	<p><i>this kind of work</i> substitutes <i>the design of the whole power station lighting...</i></p>	<p><i>...a pleasant environment and [a pleasant] atmosphere...</i></p> <p><i>...it was quite an adventure and [quite] fun as well...</i></p>	<p><i>because</i></p> <p><i>meanwhile</i></p> <p><i>and</i></p> <p><i>but</i></p> <p><i>which</i></p> <p><i>during</i></p>	<p>Physical repetition of the same item: <i>contract; permanent; lighting; design; lovely</i></p> <p>Synonyms: <i>clear x understandable; pleasant x lovely x nice</i></p> <p>Part-whole relationships: <i>building x office, churches, galleries, schools, hotels; room x furniture;</i></p> <p>Syntagmatic relations: <i>Eskom x electricity, lighting; job x employed, contract, promoted</i></p>
L2 Text 3	<p>Particular reference: <i>your</i> refers to <i>gentleman</i>; <i>them</i> refers to <i>baggers</i>; <i>it</i> refers to <i>nature</i></p> <p>Generic reference: <i>the article; the industry; the prices; the second shift</i></p> <p>Unique reference: <i>SA; Mondays; Thursdays; Fridays</i></p>	<p><i>it</i> substitutes <i>the articles on the first page</i> describe <i>sensational events.</i></p> <p><i>this</i> substitutes <i>Another problem is the society awareness and understanding of economical problems.</i></p>	<p><i>I play tennis on Mondays, [I play tennis on] Thursdays and [I] sometimes [play tennis] on Fridays.</i></p>	<p><i>and</i></p> <p><i>which</i></p> <p><i>but</i></p> <p><i>with</i></p> <p><i>perhaps</i></p> <p><i>therefore</i></p>	<p>Physical repetition of the same item: <i>the nature; people; generation; shops; lunch time; economy...</i></p> <p>Superordinate and hyponyms: <i>time x breakfast, lunch, dinner; hobbies x reading, pottery, drawing, painting</i></p> <p>Meaning opposition and contrast: <i>complicated x simple; damage x restore; abuse x use; asleep x awaked; export x import</i></p> <p>Part-whole relationships: <i>garden x flowers; newspaper x article</i></p> <p>Syntagmatic relations: <i>materials x wood, plastic; supermarket x food; food x hunger</i></p> <p>Ordered series: <i>children, grandchildren; Mondays, Thursdays, Fridays</i></p>
L2 Text 4	<p>Particular reference: <i>them</i> refers to <i>natives</i>; <i>they</i> refers to <i>the oppressors</i></p> <p>No particular reference: <i>the result; the thing; the idea; the labels</i></p> <p>Generic reference: <i>activities; pollution</i></p> <p>Unique reference: <i>Hilton; the theory about Gaja; SA</i></p>	<p><i>it</i> substitutes <i>there are millions of hungry people in our country</i></p> <p><i>it and the idea</i> substitutes <i>the theory about Gaja</i></p> <p><i>the whole process</i> substitutes <i>take from the reach and give to the poor</i></p>	<p><i>...throwing odds and [throwing] ends...</i></p> <p><i>...communicate with grass and [communicate with] trees and [communicate with] clouds...</i></p>	<p><i>by means of</i></p> <p><i>and</i></p> <p><i>of course</i></p> <p><i>in general</i></p> <p><i>although</i></p> <p><i>perhaps</i></p> <p><i>in order to</i></p>	<p>Physical repetition of the same item: <i>hungry; problem; existence; enjoyably similar</i></p> <p>Superordinate and hyponyms: <i>nature x grass, trees, clouds; flowers x magnolias</i></p> <p>Synonyms: <i>boring x tiresome; digits x numbers; weather x the sky mood</i></p> <p>Meaning opposition and contrast: <i>rich x poor; considerate x cruel; greater x lesser</i></p> <p>Part-whole relationships: <i>garden x flowers; flowers x buds; fridges x food</i></p> <p>Syntagmatic relations: <i>garden x harring; air x breath; food x eat; water x drink; weather x the sky mood x raining</i></p> <p>Ordered series: <i>second, minute, hour, day; millions, billions</i></p>

4.6 Discussion of findings

Numerous conclusions may be drawn from the study's findings. There are obvious similarities between each of the eight texts and between the L1 and L2 groups, but there are also significant differences between each author and between the two groups. The findings for each feature are discussed next.

4.6.1 Average sentence length

The second language English (L2) speakers have an overall higher average sentence length than the L1 users, owing to the fact that they join multiple sentences using the conjunction ‘and’. The first language English (L1) speakers have much shorter sentences, but they are more linguistically correct.

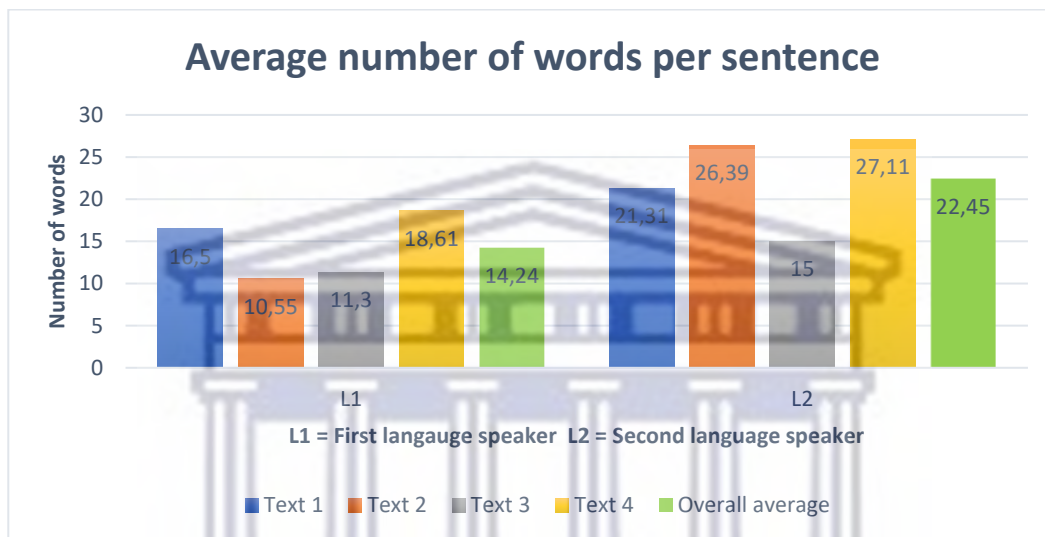


Figure 4.1: Average sentence length

4.6.2 Internal structures of sentences

The L1 speakers adhere to the grammatically correct order of parts of speech sentences – SVN (subject – verb – noun) – while two of the L2 speakers change the order a few times. The L1 speakers have an average of two clauses per sentence, while the L2 speakers’ use a range of one to six clauses per sentence. This is significant, as it clearly indicates a fundamental difference in several features of L1 and L2 speakers’ use of written English. The L1 group is consistent in their similarities, while the L2 speakers’ sentence order and clauses vary across texts.

4.6.3 End-of-sentence punctuation

All the authors use punctuation differently, depending on the genre of their texts. Three authors use only full stops; three authors use full stops, exclamation marks and question marks; and the other two authors use full stops and question marks. This feature is not quite indicative of

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

authorship in this study, because it does not indicate a difference between the two groups or between each author. The punctuation relies heavily on the content of the texts.

4.6.4 Average T-unit length

The sentences of the L1 and L2 speakers have almost similar average T-unit lengths. As stated in Section 2.10.2, the length of a T-unit determines syntactic complexity, and could be a helpful indicator of the difference between L1 and L2 writing. These results, as presented in Figure 4.2, may not seem significant on their own, but their significance becomes apparent when they are compared to the average sentence length, as discussed in Section 4.6.1. Refer to Section 4.7 for a comparison in pie graph form.

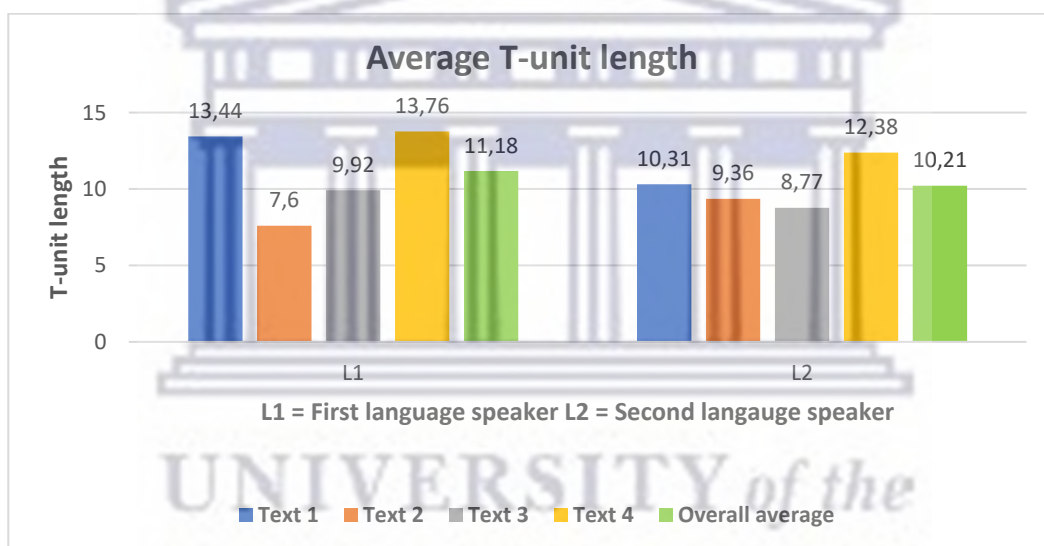


Figure 4.2 : Average T-unit length

4.6.5 Average number of T-units per sentence

The L1 authors' average number of T-units per sentence range from one to three, while those of the L2 authors range from one to four, with isolated instances of five, six and seven. This again is due to the fact that the L2 speakers join multiple sentences with the conjunction word 'and'.

4.6.6 Reference

The L1 speakers use all reference types, whereas the majority of L2 speakers use only one or two. As discussed in Section 2.10.3.1, there are four types of reference: particular reference,

no particular reference, generic reference and unique reference. Only one L2 speaker used all four. The use of all four is a sign of familiarity with a language. When an author is less well-versed in a language, they tend to use the general terms they already know rather than ‘unique’ terms; a L2 English user might write, for instance: ‘I saw *that man who works on the building* [generic reference]’, as opposed to ‘I saw the *janitor* [unique reference]’.

4.6.7 Substitution, ellipsis, conjunction and lexical cohesion

The L1 and L2 speakers used cohesion markers in very different ways. The findings show that the L2 speakers used more substitution, ellipsis and conjunction than the L1 users. In terms of number of instances, the L1 speakers actually used the same number of these features, but the L2 dataset was much smaller; hence, relative to the number of words used in total, they used the higher proportion of these features.

Also worth noting is that although L2 speakers have the higher number of conjunctions, more than half of these are the word ‘and’, used to joining multiple sentences and lists. The fact that the L2 speakers also used more ellipses than the L1 users is related to the fact that they used the conjunction ‘and’ more frequently; ‘and’ is typically followed by an ellipsis.

The L1 speakers employed greater variety in their use of conjunctions. Interestingly, because of their concern or, indeed, obsession with the past and the future, the authors of the bomb threat letters used a significant number of cohesion markers related to time.

In terms of lexical cohesion, the author of L1 Text 3 used the word *dismiss* synonymously with *kill*, which is unique, and an example of markedness. The author of L2 Text 4 used the words *the sky mood* with reference to the *weather*, which is also unique and an example of markedness. Therefore, these cohesion markers are good indicators of authorship, since their use is unique; there is a high possibility that one could attribute authorship if several texts used these terms in the same way.

Figure 4.3 shows the number of instances of substitution, ellipsis and conjunction in the four L1 texts. Figure 4.4 shows the number of the same features in the four L2 texts.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

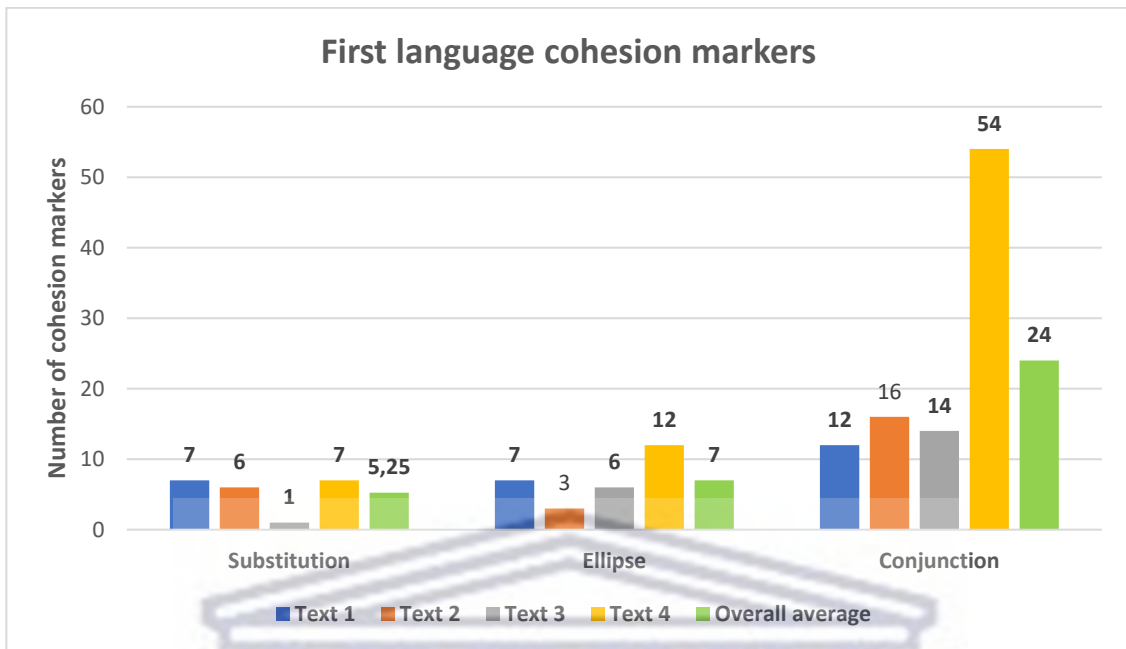


Figure 4.3: L1 cohesion markers

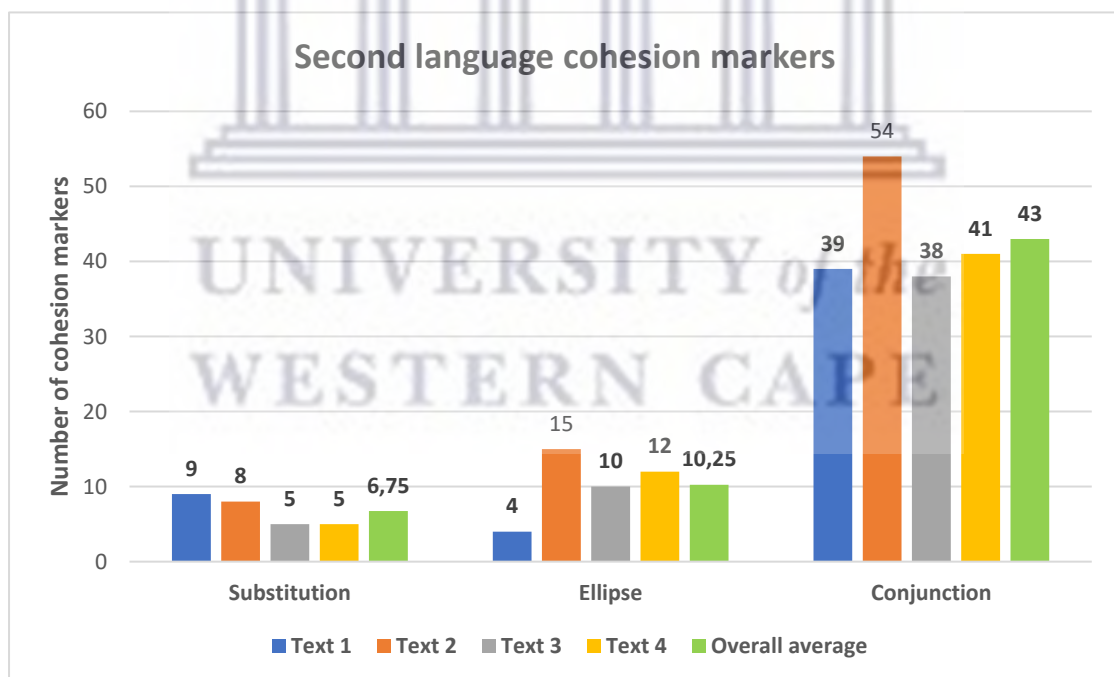


Figure 4.4: L2 cohesion markers

4.7 Statistical significance of the study

With the help of a statistician, data was extracted regarding the use of text features of interest in this study. Table 4.8 shows the relative significance of each of the these text features among the English L1 and L2 users. The table shows the probability of each feature within the text, based on the size of dataset, in other words, how unique or significant the feature is within the dataset. Statisticians prefer a probability of, and lower, than 0.1 (10%), since that shows statistical significance.

Table 4.8: Test of significance between means of L1 and L2

Indicator	Estimated probability of t-test
Words per text	0.3101
Sentences per text	0.8751
Words per sentence	0.0533
Words per t-unit	0.5818
Substitution	0.4287
Ellipse	0.3177
Conjunction	0.1261

Source: Estimated using the software EViews

The differences between the first and second language speakers' use of the investigated text features is evident in Tables 4.4 and 4.5, and is discussed in Section 4.4. From a scientific point of view, it would be ideal if the differences such as number of words per text and/or number of sentences per text could be shown to be statistically significant. Among the recognised statistical procedures to test for the statistical significance of a difference between the mean value of one group (in this case, first language speakers) and the mean value of another group (in this case, second language speakers) are t-tests, Satterthwaite-Welch t-tests, Anova F-tests and Welch F-tests. All four tests were performed in order to establish if the mean values calculated for words per text, sentences per text, words per sentence, words per t-unit, number of substitution cohesion markers, number of ellipsis cohesion markers and number of conjunction cohesion markers differed significantly in statistical terms between the two groups. For all four tests, the relevant test statistic was calculated together with the probability value. The probabilities of the four test statistics for all seven indicators were very similar and therefore only the probability for the t-test is reported in Table 4.8.

The reported probability value is an indication of 'the lowest significance level at which a null hypothesis can be rejected' (Gujarati & Porter, 2009). In layman's terms, it indicates the chance

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

of making a mistake if the null hypothesis is rejected. Statisticians usually consider probability values of 0.10 or 0.05 or 0.01 – or in other words, they are comfortable with rejecting the null hypothesis if the chance of being wrong is 10%, 5% or, even better, 1%. The null hypothesis for the test of equality of means is indicated if the mean values for both groups are equal. Therefore, the probability of 0.3101 for the first indicator in Table 4.8 means that there is a 31.01% chance of being wrong if one rejects the null hypothesis. This is too high, and the decision would be not to reject the null hypothesis and conclude that the means of the two groups are equal – or that there is not enough statistical evidence to conclude that they are not equal.

It is only the indicator of words per sentence, with a probability of 0.0533, that may be deemed statistically significant, since it gives enough statistical evidence that the two means do indeed differ enough to conclude that they are not equal. Although the probability of 0.1261, for the number of times conjunctions are used as cohesion markers exceeds the upper threshold of 10%, it provides an important indication that the means of this indicator also differs between the two groups.

As a final note: Hypothesis testing is usually done with larger samples – definitely larger than the current four per group. The reported levels of significance of 0.0533 and even 0.1261 for such a small sample indicate a remarkable difference between the two groups. If the sample size were increased, it is highly likely that the indicators ‘words per text’ and ‘ellipsis’ would also deliver statistically significant results.

4.8 Successful writing-style features

The relationship between average sentence length and average T-unit length, and the use of referencing as a cohesion marker, are two writing-style characteristics that clearly distinguish the two groups from one another. Note that L1 speakers have shorter sentences, but their average T-unit length is not significantly shorter than that of their sentences. Compare, for example, the following ratios of T-unit lengths to sentence lengths for L1 speakers: 13,44: 16,5; 7,6: 10,55; 9,92:11,3; and 13,76: 18,0 with the average ratios for T-unit length to sentence length of L2 speakers: 10,31: 21,31; 9,36: a 26,39; 8,77: 15; and 12,38: 27,11. This indicates that the T-units make up roughly half of the L2 speakers’ sentences. This is due to the fact that their sentences are less complex and contain multiple occurrences of the conjunction ‘and’.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

Using the average T-unit length and average sentence length of each group, the graphs shown in Figure 4.5 were made. Here one can see how much of each sentence is a T-unit, and therefore how complex the sentence is. The L1 speakers' sentences are 32.7% more complex than the L2 speakers'.

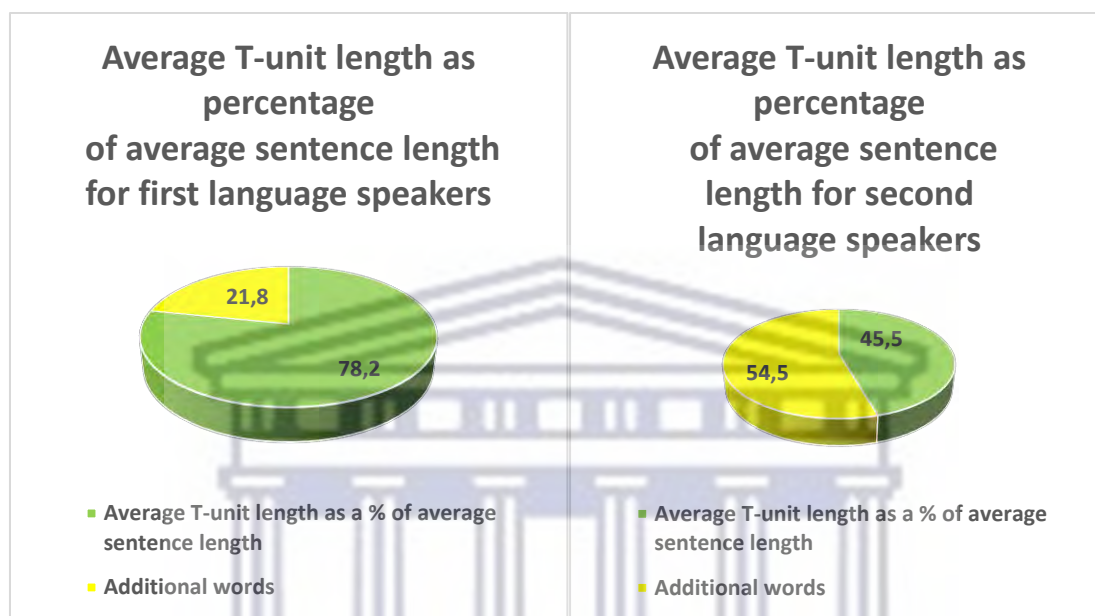


Figure 4.5: Comparison between L1 and L2 average T-unit length relative to average sentence length

The cohesion markers discussed indicate some areas of markedness or uniqueness among the authors. The examples shown Table 4.6 and 4.7 show the authors' preferences for certain cohesion markers. The authors all use reference, substitution, ellipsis and lexical cohesion differently – from the examples in Table 4.6 and 4.7, it may be deduced that, if more texts by the same authors were available, more similarities would be seen among the texts.

The use of lexical cohesion in the texts differs based on the topic, but also on the author, since every author uses them differently. For example, the author of L1 Text 1 uses the lowest number of lexical cohesion markers, with the part-whole relations given indicating the detail with which the author explains the parts of a bomb. The author of L1 Text 3 uses the word 'dismiss' synonymously with 'kill', as indicated in Section 4.6.7.

The author of L1 Text 4 has a tendency to use a lot of opposition and contrast in terms of meanings of words, as well as syntagmatic relations and synonyms. The analysis of the use of superordinates and hyponyms reveals that the author of L2 Text 4 describes *feelings* very often.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

This author uses the most forms of lexical cohesion, with a particular example of markedness apparent in the term *the sky mood*, used synonymously with *the weather*.

The framework was effective, but in the future, it would be preferable to focus on fewer writing-style features. The suggested writing style characteristics for the method are average sentence length relative to T-unit length and cohesion markers. The employed classification method, *WordSmith Tools*, was useful for quantifying the results, but a tool that automatically tags these characteristics for the researcher would be ideal. Most of the tagging had to be done by hand using coloured pens, as there is no available software that tags these markers. The analysis took quite a long time because of this.



CHAPTER 5: CONCLUSION AND RECOMMENDATION

5.1 Answers to the research questions

The findings of this research have answered the research questions with varying degrees of comprehensiveness.

The first research question was: ‘To what extent is a standardised method for authorship identification/attribution possible, using a certain combination of writing-style features?’ Based on the literature review and the findings, a conclusive answer cannot be given. There is a need for several texts by the same author, enough in number and length to yield statistically significant results. The hypothesis arises that a standardised method for authorship attribution is possible if a bigger corpus is created with texts of similar genres.

The second research question was: ‘What are the available writing style features, and to what extent are they effective for authorship identification/attribution of L1 and L2 English texts?’ As discussed in Chapter Two, writing-style features may be lexical, syntactic, structural and content-specific. This study was concerned only with syntactic writing-style features, as the research shows that most studies have moved away from considerations of the others, finding the syntactic features to be more telling and better indicators of writing complexity than the other three feature types. Not all the writing-style features chosen for this study delivered significant outcomes; however, two stood out as indicative of L1 and L2 writing and individual writing choices – T-unit length relative to average sentence length, and use of cohesion markers.

The third research question was: ‘Which classification techniques are effective for authorship identification/attribution of L1 and L2 English texts?’ As seen from the literature review, there are multiple techniques available, such as CUSUM, n-grams, neural networks and Markov’s chains, but none of these would have been appropriate for this study, as a great part of the analysis relied on a qualitative approach, and none of the classification techniques cater especially for the features identified. Here, again, the issue was that the dataset was too small. *WordSmith* was helpful for confirming the analysis done by hand, in that it created concordances in which words appeared in their context, together with the number of

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

appearances. If a bigger dataset were acquired, the statistical significance t-test discussed in Section 4.7 would be the preferred classification technique to use.

Finally, the last research question was: ‘Can the proposed standardised method be deployed on online texts?’ This study has shown that the proposed method can be deployed successfully on texts of various sizes, including shorter online messages such as text messages and e-mails. The analysis of the features chosen would reveal markers indicative of authorship. The classification technique might be troublesome if one did not have a large dataset; however, even if the results cannot be quantified using a classification technique, the identified features and the method of analysis may be used to identify authors. It would, of course, be helpful to have several texts for each author, but the methodology will work even if only one text is available for each author, as was the case in this study.

5.2 Summary of findings

A mixed method approach was used in the analysis of the texts. The qualitative analysis consisted of tagging the identified writing-style features: (i) end-of-sentence punctuation, (ii) internal structure of sentences, (iii) average sentence length, (iv) average T-unit length, (v) number of T-units per sentence, (vi) use of referencing, (vii) use of substitution, (viii) use of ellipses, (ix) use of conjunctions and (x) use of lexical cohesion. All of this was done by hand. The quantitative analysis involved creating a corpus of the eight texts on *WordSmith Tools*, which allowed the researcher to create concordances, with *WordSmith* showing the tagged linguistic features in their context and the number of times each feature was used. This was quite tedious work, as the software could not show the specific instances of cohesion, sentence length or T-units, etc., but only the number of times a word occurred, which the researcher then had to search for in the concordance. Despite these drawbacks, findings could still be made, as recorded in Tables 4.6 and 4.7.

The numbers shown on the tables were used to generate graphs, which showed the differences between the use of the linguistic features by all eight authors, as well as between the two separate groups – L1 and L2 users. The findings from the graphs indicate that no author made the same choices as another; each author use all the writing-style features differently. However, certain writing-style features were better indicators of the differences between the two groups than others: the average T-unit length relative to the average sentence length was a particularly

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

valuable indicator, as shown in Figure 4.5, and the cohesion markers, as discussed in Section 4.6.7 and shown in Tables 4.6 and 4.7. A statistician also provided insight into the statistical significance of the findings (see Section 4.7), performing four different statistical analyses with the amounts generated for Table 4.4. The only markers that were statistically significant based on the probability test were words per sentence and use of conjunctions. The statistician recommended bigger datasets for future study, as this would probably show statistical significance for number of words per text and use ellipsis as well.

Thus, the combined analyses and findings thereof showed that the following chosen writing-style features may be used for authorship attribution: average sentence length, or words per sentence, average T-unit length relative to average sentence length, references, substitution, ellipses, conjunctions and lexical cohesion markers. Future analysis will be most successful if the mixed methods used in the current study are employed, with qualitative analysis used to identify the features of interest, and quantitative methods used – with the help of a statistician – to generate the numbers of instances of each marker. Lastly, the statistician would have to conduct a probability test.

5.3 Limitations of the study

This study was mostly successful, as answers to the research questions could be acquired, despite the following limitations:

- This study would have yielded more conclusive results if the researcher had had access to multiple texts by the same author. With more texts from which to extract data, it would have been possible to compare the instances of writing-style features used in different texts, and to conclude whether certain authors always use the same features in similar ways.
- The length of the texts varied greatly between the L1 and L2 group, which made it harder to compare quantities. This limitation was overcome by noting not only the number of times a text feature was used, but by considering the occurrence in relation to total corpus size for each group.
- There is no software available to tag the writing-style features under investigation. The reliance on manual tagging opened up the possibility of errors creeping in during the qualitative phase of analysis.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

- A bigger corpus/dataset is needed, so that the statistical significance of features would be below 10%.

5.4 Recommendations for further study

This research study aimed to offer a framework for linguistic analysis of texts that would allow the creation of a possible standardised method for authorship attribution. The framework and method used in this study were helpful, despite their limitations, and may serve as a foundation for future research. Furthermore, further research is warranted and necessary to validate the framework. Further research should also be done in other languages, with the aim of testing the method more vigorously. The research should also be conducted on texts written by English L3, L4 and L5 texts, since, as noted in Section 1.4, South Africa is a particularly linguistically diverse country, in which many people use English as their third or even fourth and fifth language.

Future research should be done on texts by the same author. The method has been shown to work on texts by different authors, revealing the uniqueness of their writing styles. The research should also be conducted on shorter texts, such as e-mails. An interesting discovery was made regarding the cohesion markers used in bomb threats – most of the markers were indicative of time, as most bomb threats state exactly when and where the writer will do something (or has done something). These authors, too, tend to be highly concerned about things that, in their view, have gone wrong in society in the past or are currently going wrong, so that time references are common. It would be interesting to compare texts of this nature and investigate whether all bomb threat letters share this feature, and whether time markers might be used to provide a framework for features in bomb threats specifically.

The framework and proposed method has applicability to cases of plagiarism and should be used for this purpose, since the method has been shown to work well with longer texts and essays. Further research could use fewer writing-style features, as recommended under Section 4.8, to conduct authorship profiling of age, gender and educational level.

Finally, further research should be conducted to create software that can tag linguistic features identified and generate quantitative output, while also comparing texts, so that results may be generated more quickly. A component of this software should be that it contains built-in corpuses of various genres of texts, varying in length and categorised by language proficiency.

5.5 Conclusion

This study has presented comprehensive data on the field of forensic linguistics through an analysis of eight key texts. The results have provided clear answers to the research questions. The salient writing-style features were identified and the classification techniques used for each feature were discussed, along how the classifications may be improved.

The results show that a standardised method for authorship attribution is possible, since the method used in this study revealed clear distinctions between the use of certain linguistic features in all eight authors, and between the body of L1 users and L2 users. It is notable that among the small corpus of work studied, all the authors used the writing-style features differently to one another; no single author used any feature in the same way as another. However, this may be seen as a feature of the small sample size, and it may be hypothesised that a larger corpus of work would reveal similarities between texts. This would be useful for verifying the use of unique linguistic features by particular authors. In answer to the last research question ('Is the proposed standardised method efficient enough to be deployed on online texts?') the study confirmed that this method works on different types of texts, including online texts. The shorter texts were ultimately easier to analyse, since they required less time to tag.

Broadly speaking, the study set out to determine where authorship analysis is situated in the field of forensic linguistics, and what the role of the forensic linguist is in court. This aspect was investigated in order to draw conclusions on whether forensic linguists can in fact be considered expert witnesses, and what is needed in order for them to be recognised as such. It is obvious that courts all over the world require evidence that cannot be called into question; for this reason, a quantifiable method is necessary for authorship attribution. This led to an investigation of frameworks and methods that are permissible in court; it was found that an acceptable method exists that gives data in the form of numerical/statistical outputs. This method is called stylometry, and it exists in the discipline of linguistic forensics.

The researcher then set out to investigate the writing-style features, and a classification system for them, that could most successfully be used to determine authorship. The identified features were used to create a conceptual framework as a basis for the proposed authorship attribution method. The framework, adapted from Zheng et al. (2006), was explained in terms of the various steps that have to be taken, starting with data collection, moving to feature extraction, then method generation and finally authorship attribution. The texts analysed generated

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

workable findings and proved that seven of the initial ten writing-style features can be used for authorship attribution: (i) average sentence length, (ii) average T-unit length, (iii) use of references, (iv) use of substitution, (v) use of ellipses, (vi) use of conjunctions and (vii) lexical cohesion.

The topic chosen for this research project falls under the broader subject field of forensic linguistics, which is concerned with ‘the analysis of language that relates to the law, either as evidence or as legal discourse’ (Olsson & Luchjenbroers, 2014, p. 1). The purpose of this study was to provide a standardised method for authorship attribution in L1/L2 English texts. The research questions were answered based on a detailed and systematic literature review, in conjunction with a mixed method analysis of eight L1/L2 English texts, sampled for the purpose of the research.

Four research questions were stipulated to guide the researcher towards a possible standardised method for authorship attribution. The answers to the questions were gained by following the methodology set out in Chapter Three. In answering the questions, the research fulfilled its objectives as set on in Section 1.5, which were:

- i. To investigate to what extent a standardised method for authorship attribution is possible;
- ii. To identify the available writing style features, and the extent of their effectiveness for authorship identification/attribution of L1 and L2 English texts;
- iii. To identify the classification techniques and the extent of their effectiveness for authorship identification/attribution of L1 and L2 English texts;
- iv. To contribute to the study of forensic linguistics with particular reference authorship attribution.

Despite the identified limitations of the study, this research provides a foundation for the analysis of various texts where authorship needs to be attributed. It was possible to create a method for authorship attribution that is applicable to various text types. This method could be applied in many different investigations, such as those to do with plagiarism, defamation, hate mail, threatening texts and suicide notes.

To the best of this author’s knowledge, this research is the first of its kind in South Africa. It is hoped that the findings of this study will form the foundation for, or at least a useful

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

contribution to, many new topics of research in the general area of authorship attribution. The study has contributed to the field of applied forensic linguistics. It is particularly hoped that it will inspire the development of an even more comprehensive and detailed framework for providing objective evidence about authorship in forensic cases.



REFERENCES

- Abbasi, A. and Chen, H. 2005. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5), pp. 67a75.
- Afroz, S., Islam, A.C., Stolerman, A., Greenstadt, R. and McCoy, D. 2014, May. 'Doppelgänger finder: Taking stylometry to the underground', *2014 IEEE Symposium on Security and Privacy* (pp. 212–226). IEEE.
- Altum, J.C. 2003. Anti-abortion extremism: The army of God. *Chrestomathy: annual review of undergraduate research at the College of Charleston*, 2, pp.1–12.
- Angouri, J. 2010. Quantitative, qualitative or both? Combining methods in linguistic research. In: Litosseliti (ed.). *Research methods in linguistics*. London: Continuum International Publishing Group. pp. 29–45.
- Argamon, S., Šarić, M. and Stein, S.S. 2003, August. 'Style mining of electronic messages for multiple authorship discrimination: First results', *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 475–480).
- Argamon, S. and Levitan, S. 2005, June. 'Measuring the usefulness of the function words for authorship attribution', *Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing* (pp. 1–3).
- Babbie, F. and Mouton, J. 1998. *The practice of social research*. Oxford University Press.
- Barber, A. 2004. Idiolects. In: E.N. Zalta (ed.), *The Stanford encyclopedia of philosophy*: Winter 2004 edition.
- Barry, K. and Luna, K. 2012. *Stylometry for online forums*. Stanford University.
- Battistella, E. 1995. Jakobson and Chomsky on markedness. In Haji ová, E., Miroslav Ervenka, M., Le ka, O., and Sgall, P. (eds.). *Prague Linguistic Circle Papers: Language, Arts and Disciplines*. p. 55–71. John Benjamins Publishing.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

- Biber, D., Conrad, S. and Reppen, R. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
- Blackwell, S. 2012. History of forensic linguistics. *The encyclopedia of applied linguistics*.
- Brennan, M, Afroz, S and Greenstadt, R. 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security*, 15(3).
- Broeders, T. 2001, July. 'Forensic speech and audio analysis forensic linguistics 1998–2001', *Proceedings 13th INTERPOL Forensic Science Symposium, Lyon, France D* (Vol. 2, pp. 54–84).
- Brown, J. 2004. Research methods for applied linguistics, scope, characteristics and standards. In A. Davies, and C. Elder (Eds.). *The handbook of applied linguistics*. Blackwell Publishing. 476–500.
- Brown, G. and Yule, G. 1983. *Discourse analysis*. Cambridge University Press.
- Brown-Jackson, M. 2013. *How linguistic analysis helped unmask Robert Galbraith as J.K Rowling*. Geekosystem. Available at: <https://www.themarysue.com/linguistic-tool-rowling/> (accessed on: 20 February 2023).
- Carney, T.R. 2012. 'n Forensies-semantiese beskouing van die woordgebruik 'onkoste' in die hofsaak Commissioner for South African Revenue Service vs. Labat Africa Limited. *South African Linguistics and Applied Language Studies*, 30(4): 487–496.
- Carstens, W.A.M. 2016. *Afrikaanse tekslinguistiek: 'n Inleiding*. (10th ed.). J.L. van Schaik Uitgewers.
- Carstens, W.A.M. and Van de Poel, K. 2012. *Teksredaksie*. African Sun Media.
- Cavnar, W.B. and Trenkle, J.M. 1994, April. 'N-gram-based text categorization', *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval* (Vol. 161175).

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

- Chaski, C.E. 2001. Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8, pp.1–65.
- Chaski, C.E. 2005. Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4(1).
- Chaski, C.E. 2007. The keyboard dilemma and authorship identification. *Advances in Digital Forensics III*, pp. 133–146.
- Chaski, C.E. 2007. ALIAS Technologies LLC. Institute for Linguistic Evidence.
- Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Available at: https://babel.ucsc.edu/~hank/aspects_ch3.pdf (accessed on: 19 January 2023).
- Church–Fleischmann, A. 2020. *n Korpusegebaseerde ondersoek na kohesiepatrone as moontlike stilistiese kenmerk van outeurstyl* (Doctoral dissertation, University of the Free State).
- Corney, M. W. 2003. *Analysing e-mail text authorship for forensic purposes* (Masters dissertation, Queensland University of Technology).
- Coulthard, M. 2004. Author identification, idiolect and linguistic uniqueness. *Applied Linguistics*, 25(4), pp. 431–447.
- Coulthard, M. 2010. Experts and opinions: In my opinion. In: *The Routledge handbook of forensic linguistics*. Routledge.
- Coulthard, M. and Johnson, A. 2007. *An introduction to forensic linguistics: Language in evidence*. Routledge.
- Coulthard, M. and Johnson, A. 2010. *The Routledge handbook of forensic linguistics*. Routledge.
- Coulthard, M., Grant, T. and Kredens, K. 2011. Forensic linguistics. In R. Wodak, B. Johnstone, B. and P. Kerswill (eds.) *The SAGE handbook of sociolinguistics*. Sage, pp. 529–544.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

- Coulthard, M., Johnson, A. and Wright, D. 2016. *An introduction to forensic linguistics*. (2nd ed.). Taylor and Francis.
- Crankshaw, R. 2012. The validity of the linguistic fingerprint in forensic investigation. *Diffusion* 5(2). University of Central Lancashire.
- Crystal, D. 2008. *Txtng: The gr8 db8*. Oxford: Oxford University Press.
- De Klerk, W.J. 1978. *Inleiding tot die semantiek*. Butterworth.
- De Stadler, L.G. 1989. *Afrikaanse semantiek*. Southern Uitgewers.
- 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.
- De Vries, A. and Docrat, Z. 2019. Multilingualism in the South African legal system. *New Frontiers in forensic linguistics: Themes and perspectives in language and law in Africa and beyond*, p. 89.
- Docrat, Z. 2018. *The role of African languages in the South African legal system: Towards a transformative agenda*. [Doctorate thesis]. Rhodes University.
- Docrat, Z. 2022. A Review of linguistic qualifications and training for legal professionals and judicial officers: A call for linguistic equality in South Africa's legal profession. *International Journal for the Semiotics of Law/Revue Internationale de Sémiotique Juridique*, 35(5), pp. 1711–1731.
- Docrat, Z., Kaschula, R.H. and Ralarala, M.K. eds. 2021. *A handbook on legal languages and the quest for linguistic equality in South Africa and beyond (Vol. 3)*. African Sun Media.
- Dörnyei, Z. 2007. *Research methods in applied linguistics*. Oxford University Press.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

- Egins, S. 2004. *An introduction to systemic functional linguistics*. Continuum International Publishing Group.
- El, S.E.M. and Kassou, I. 2014. Authorship analysis studies: A survey. *International Journal of Computer Applications*, 86(12).
- Gaies, S. 1980. T-unit analysis in second-language research: Applications, problems, and limitations. *TESOL Quarterly*, 14(1), pp. 53–60.
- Galal, S. 2022. Distribution of languages spoken inside and outside of households South Africa 2018. *Statista*. Available at: <https://www.statista.com/statistics/1114302/distribution-of-languages-spoken-inside-and-outside-of-households-in-south-africa/> (accessed on: 27 September 2023).
- Gales, T. 2015. The stance of stalking: A corpus-based analysis of grammatical markers of stance in threatening communications. *Corpora*, 10(2), pp. 171–200.
- Gavaldà-Ferré, N. 2012. 'The study of inter- and intra-speaker variation towards an index of idiolectal similitude' *Proceedings of the International Association of Forensic Linguists' Tenth Biennial Conference 11 to 14 July 2011*. Birmingham: Aston University.
- Grant, T. 2007. Quantifying evidence in forensic authorship analysis. *International Journal of Speech, Language and the Law*, 14(1).
- Grant, T. 2008. Approaching questions in forensic authorship analysis. *Dimensions of Forensic Linguistics*, 5, p. 215.
- Grant, T. 2010. Txt 4n6: Idiolect free authorship analysis? In: *The Routledge handbook of forensic linguistics*. Routledge
- Grieve, J.W. 2005. Quantitative authorship attribution: A history and an evaluation of techniques. [Master's thesis]. Simon Fraser University.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

Grieve, J. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 223, pp. 251–270.

Guillén Nieto, V., Vargas Sierra, C., Pardiño Juan, M., Martínez Barco, P., and Suárez Cueto, A. 2008. Exploring state-of-the-art software for forensic authorship identification. *International Journal of English Studies* 8(1):1–28.

Gujarati, D.N. and Porter, D.C. 2009. *Essentials of econometrics* (4th ed.). McGraw-Hill Publications.

Halliday, M.A.K. 1975. *Learning how to mean: Explorations in the development of language*. Edward Arnold.

Halliday, M.A.K. and Hasan, R. 1976. *Cohesion in English*. Longman.

Hasan, R. and Cloran, C. 1990. A sociolinguistic interpretation of everyday talk between mothers and children. In M. G. Halliday Red., *Learning, keeping and using language: Selected papers from the 8th world congress of applied linguistics*, Sydney, 16–21 Augustus 1987. John Benjamins Publishing Company. 67–100.

Hirano, K. 1989. Research on T-unit measures in ESL. *Bull. Joetsu University Education*, 8(2): pp. 67–77.

Hoey, M. 1991. *Patterns of lexis in text*. Oxford University Press.

Hofman, J. 2006. Electronic evidence in criminal cases. *South African Journal of Criminal Justice*, 19(3), pp. 257–275.

Holmes, D.I. 1994. Authorship attribution. *Computers and Humanities*, 28, pp. 87–101.

Holmes, D.I. 1998. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3), pp. 111–117

Hubbard, E.H. 1994. Errors in court: A forensic application of error analysis. *SA Journal of Linguistics – Supplement* 20, pp. 3–16.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

Hubbard, E.H. 1995. Linguistic fingerprinting: A case study in forensic linguistics. *SA Journal of Linguistics – Supplement 26*, pp. 55–72.

Hubbard, E.H. 2009. ‘Stylometric and error analysis in the context of a style shift in abusive e-mail texts’, *9th Biennial Conference on Forensic Linguistics/Language and Law*, VU University, Amsterdam.

Hunt, K.W. 1965. *Grammatical structures written at three grade levels*. NCTE Research Report, No. 3. Urbana, IL: The National Council of Teachers of English.

Iqbal, F., Hadjidj, R., Fung, B.C. and Debbabi, M. 2008. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *Digital Investigation*, 5, pp. S42–S51.

Iqbal, F., Khan, L.A., Fung, B.C. and Debbabi, M. 2010, March. ‘E-mail authorship verification for forensic investigation’, *Proceedings of the 2010 ACM Symposium on Applied Computing*, pp. 1591–1598.

Ishihara, S. 2011. ‘A forensic authorship classification in SMS messages: A likelihood ratio based approach using N-gram’, *Proceedings of Australasian Language Technology Association Workshop*, pp. 47–56

Jackson, A.R.W. and Jackson, J.M. 2004. *Forensic science*. Pearson Education Limited.

Jakobson, R. 1963. Implications of language universals for linguistics. *Universals of language*, 208, p. 219.

Johnson, A. and Coulthard, M. 2010. Current debates in forensic linguistics. In: *The Routledge handbook of forensic linguistics*. Routledge.

Johnson, K. and Johnson, H (eds.). 1999. Macro/microlinguistics. In: *Encyclopaedic dictionary of applied linguistics*. Blackwell Publishing.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

- Juola, P. 2006. Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3), pp. 233–334
- Kacmarcik, G. and Gamon, M. 2006, July. ‘Obfuscating document stylometry to preserve author anonymity’, *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pp. 444–451.
- Kean, M.L. 1975. The theory of markedness in generative grammar. [Doctoral thesis]. Massachusetts Institute of Technology.
- Khmelev, D en Tweedie, F.J. 2001. Using Markov chains for identification of writers. *Literary and Linguistic Computing*, 16(3), pp. 299–307.
- Klopper, R. 2009. The case for cyber forensic linguistics. *Alternation*, 16(1), pp. 261–294
- Koppel, M and Schler, J. 2004. ‘Authorship verification as a one-class classification problem’, *Proceedings of the Twenty-First International Conference on Machine Learning (ICML)*, pp. 489–495.
- Koppel, M., Schler, J. and Argamon, S. 2009. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1), pp.9-26.
- Kotzé, E.F. 2007. Die vangnet van die woord: Forensies-linguistiese getuienis in ’n lastersaak. *Southern African Linguistics and Applied Language Studies*, 25(3), pp. 385–399.
- Kotzé, E.F. 2010. Author identification from opposing perspectives in forensic linguistics. *Southern Africa Linguistics and Applied Language Studies* 28(2), pp. 185–197.
- Larsen-Freeman, D. and Strom, V. 1977. The construction of a second language acquisition index of development. *Language Learning*, 27(1), pp.123–134.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

- Leonard, R.A. 2006. Forensic linguistics: Applying the scientific principles of language analysis to issues of the law. *International Journal of the Humanities*, 3(7).
- Lombard, E. and Carney, T.R. 2011. Die wenslikheid van Afrikaans as vaktaal vir regstudente. *Potchefstroomse Elektroniese Regsjoernaal*, 14 (1), pp. 164–187.
- Luyckx, K. and Daelemans, W. 2011. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, 26(1), pp. 35–55.
- Matthiessen, C. M. 2007. The ‘architecture’ of language according to systemic functional theory: developments since the 1970s. In R. Hasan, C. Matthiessen, and J. J. Webster (eds.). *Continuing discourse on language: A functional perspective*. Equinox. pp 505–561.
- McEnery, T., Xiao, R. and Tono, Y. 2006. *Compus-based language studies: An advanced resource book*. Routledge.
- McLeod, N. and Grant, T. 2012. ‘Whose tweet? Authorship analysis of micro-blogs and other short-form messages’, *Proceedings of the International Association of Forensic Linguists’ Tenth Biennial Conference 11 to 14 July 2011*. Birmingham, Aston University.
- McMenamin, G. R. 2002. *Forensic linguistics: Advances in forensic stylistics*. CRC Press.
- McMenanim, G.R. 2010. Theory and practice of forensic stylistics. In: *The Routledge handbook of forensic linguistics*. Routledge.
- Mendenhall, T.C. 1887. The characteristic curves of composition. *Science*, 9(214), pp. 237–249.
- Michell, C.S. 2013. *Investigating the use of forensic stylistic and stylometric techniques in the analyses of authorship on a publicly accessible social networking site (Facebook)* [Doctoral dissertation] University of South Africa.
- Mikros, G.K. 2012. Authorship attribution and gender identification in Greek blogs. *Methods and Applications of Quantitative Linguistics*, 21, pp. 21–32.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

- Mitchell, E. 2008. The case for forensic linguistics. *BBC News*, 8 September.
- Moeketsi, R.H. 1999. Of African languages and forensic linguistics: The South African multilingual and multicultural criminal courtroom. *Dissertation Abstracts International, A: The Humanities and Social Sciences*, 59, pp. 4125-4126.
- Mosteller, F. and Wallace, D.I. 1964. *Interference and disputed authorship: The federalist*. Stanford University Center for the Study of Language and Information.
- Mouton, J., and Marais, H. 1990. *Basiese begrippe: metodologie van die geesteswetenskappe*. RGN Uitgewers.
- National Research Council of the National Academics. 2009. *Strengthening forensic science in the United States: A path forward*.
- Nini, A. n.d. Forensic authorship analysis. Available at: <https://andreanini.com/forensic-authorship-analysis/> (accessed on: 16 March 2023).
- Nini, A. 2015. *Authorship profiling in a forensic context* [Doctoral dissertation]. Aston University.
- Nini, A. and Grant, T. 2013. Bridging the between stylistic and cognitive approaches to authorship analysis using systemic functional linguistics and multidimensional analysis. *The International Journal of Speech, Language and the Law*, 202, pp. 173–202.
- Olsson, J. 2004. *Forensic linguistics: An introduction to language, crime and the law*. Continuum International Publishing Group.
- Olsson, J. 2008. *Forensic Linguistics*. (2nd ed.). Continuum International Publishing Group.
- OpenAI. 2022. *Introducing ChatGPT*. Available at: <https://openai.com/blog/chatgpt> (accessed on: 1 May 2023).
- Perkins, K. 1980. Using objective methods of attained writing proficiency to discriminate among holistic evaluations. *TESOL Quarterly*, 14(1), pp. 61–69.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

- Perkins, R. and Grant, T. 2013. Forensic linguistics. In J.A. Siegel, and P.J. Saukko (eds.) *Encyclopaedia of forensic sciences* (2nd ed.). Academic Press, pp. 174–177.
- Philbrick, F.A. 1949. *Language and the law: The semantics of forensic English*. The Macmillan Company.
- Ralarala, M.K. 2012. A compromise of rights, rights of language and rights to a language in Eugene Terre'Blanche's (ET) trial within a trial: Evidence lost in translation. *Stellenbosch Papers in Linguistics*, 41, pp. 55–70.
- Ralarala, M.K. 2013. 'Meaning rests in people not in words': Linguistic and cultural challenges in a diverse South African legal system. Van Schaik.
- Ralarala, M.K. 2014. Transpreters' translations of complainants' narratives as evidence: whose version goes to court? *The Translator*, 20(3), pp. 377–395.
- Ralarala, M.K. 2016. An analysis of critical 'voices' and 'styles' in transpreters' translations of complainants' narratives. *Translation and Translanguaging in Multilingual Contexts*, 2(1), pp. 142–166.
- Ralarala, M., Kaschula, R. and Heydon, G. (eds). 2019. *New frontiers in forensic linguistics: themes and perspectives in language and law in Africa and beyond*. African Sun Media.
- Ralarala, M.K., Kaschula, R.H. and Heydon, G. eds. 2022. *Language and the law: Global perspectives in forensic linguistics from Africa and beyond* (Vol. 3). African Sun Media.
- Rao, J.R. and Rohatgi, P. 2000, August. 'Can pseudonymity really guarantee privacy?' *USENIX Security Symposium*, pp. 85–96.
- Reddy, V. and Potgieter, C. 2006. 'Real men stand up for the truth': Discursive meanings in the Jacob Zuma rape trial. *Southern African Linguistics and Applied Language Studies*, 24(4), pp. 511–521.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

- Regenesys Business School. 2020. *The Fifth Industrial Revolution (5IR) and how it will change the business landscape*. Available at: <https://www.regenesys.net/reginsights/the-fifth-industrial-revolution-5ir/> (accessed on: 1 May 2023).
- Rudman, J. 1998. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31, 351–365.
- Sanderson, P. 2007. Linguistic analysis of competing trademarks. *Language Matters*, 38(1), pp. 132–149.
- Schleppegrell, M. 2013. Systemic functional linguistics. In J. P. Gee and M. Handford (eds.), *The Routledge handbook of discourse analysis*. Taylor and Francis Ltd. pp. 21–34.
- Schulstad, I., Boga, M., Jordan, C., Pally, K., Monaco, J., DeStefano, R., Stewart, J. and Tappert, C. 2012. 'Evaluation of a stylometry system on various length portions of books', *Proceedings of Student-Faculty Research Day, CSIS, Pace University*, pp. 51–58.
- Scott, M. 2021. *Wordsmith tools version 8*. Liverpool: Lexical Analysis Software.
- SCSA (Supreme Court of South Africa). 1989. 'Report compiled on language usage of Dr Victor Bran as requested by Advocate A.P. Bezuidenhout.' Case 156/89, Exhibit 32.
- Smith, M.W.A. 1983. Recent experience and new developments of methods for the determination of authorship. *Association for Literary and Linguistic Computing Bulletin*, 11, pp. 73–82.
- Soiferman, L.K. 2010. *Compare and contrast, inductive and deductive research approaches*. University of Manitoba.
- Solan, L.M. 2010. The forensic linguist: The expert linguist meets the adversarial system. In: *The Routledge handbook of forensic linguistics*. Routledge
- Somers, H. 2008. *Stylometry and Authorship*. [Powerpoint]. University of Manchester: School of Computer Science. Available at: <https://personalpages.manchester.ac.uk/staff/harold.somers/LELA30922/>

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

- Stamatatos, E, Fakotakis, N. and Kokkinakis, G. 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35, pp. 193–214.
- Stamatatos, E. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3), pp. 538–556.
- Stotsky, S. 1983. Types of lexical cohesion in expository writing: implications for developing the vocabulary of academic discourse. *College Composition and Communication* 34, pp. 430–446.
- Svartvik, J. 1968. *The Evans statements*. University of Goteburg.
- Taylor, D.C. 1998. Addressing the insane language of law. *Tydskrif vir Hedendaagse Romeins-Hollandse Reg*, 61(1), pp. 668–677.
- Thetela, P.H. 2002. Sex discourses and gender constructions in Southern Sotho: A case study of police interviews of rape/sexual assault victims. *Southern African Linguistics and Applied Language Studies*, 20(3), pp. 177–189.
- Thiart, L. 2015. Outeuridentifikasie: 'n Forensies-taalkundige ondersoek na SMS-taal in Afrikaans (Masters dissertation, University of Pretoria).
- Tkacukova, T. 2019. Forensic linguistics and language and the law. In: Schmitt, N. and H. Rodgers, M. P. *An introduction to applied linguistics* (3rd ed.). Routledge.
- Turell, M.T. 2004. Textual kidnapping revisited: The case of plagiarism in literary translation. *International Journal of Speech, Language and the Law*, 11(1), pp.1–26.
- Turell, M.T. 2008. An introduction to forensic linguistics: Language in evidence. *Atlantis*, pp. 155-160.
- Turell, M.T. 2010. The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *International Journal of Speech, Language and the Law*, 17(2).
- Van den Berg, K. 2019. A case of crying wolf? In: *New frontiers in forensic linguistics: Themes and perspectives in language and law in Africa and beyond*, p. 301.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

- Van den Berg, K. and Surmon, M. 2019. The act of threatening: Applying speech act theory to threat texts. *New frontiers in forensic linguistics: themes and perspectives in language and law in Africa and beyond*, p. 255.
- Watney, M. 2009. Admissibility of electronic evidence in criminal proceedings: An outline of the South African legal position. *Journal of Information, Law and Technology*, 2009 (1), pp. 1–13.
- Wei, L.(ed). 2011. From pedagogical practice to critical enquiry: an introduction to applied linguistics. In: *The Routledge applied linguistics reader*. New York: Routledge.
- Wybenga, D.M. 1988. *Diskoersanalise en stilistiek*. Pretoria: Serva.
- Zax, D. 2014. How did computers uncover JK Rowling’s pseudonym? *Smithsonian Magazine*. Available online: <https://www.smithsonianmag.com/science-nature/how-did-computers-uncover-jk-rowlings-pseudonym-180949824/> (accessed on: 21 April 2023).
- Zechner, N. 2013. The past, present and future of text classification. *Methods*, 1, p. 4.
- Zhang, C., Wu, X., Niu, Z., and Ding, W. 2014. Authorship identification from unstructured texts. *Knowledge-Based Systems*, 66, pp 99–111.
- Zhao, Y. and Zobel, J. 2007, January. ‘Searching with style: Authorship attribution in classic literature’, *Proceedings of the Thirtieth Australasian Conference on Computer Science* 62, pp. 59–68.
- Zheng, R., Li, J., Huang, Z. and Chen, H. 2006. A framework for authorship analysis of online messages: Writing-style features and techniques, *Journal of the American Society for Information Science and Technology* 57(3) pp. 378–393.

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

Appendix 1

Text of case study: Army of God

THE BOMBING'S IN SANDY SPRING'S AND MIDTOWN WHERE CARRIED OUT BY THE UNITS OF THE ARMY OF GOD¹

YOU MAY CONFIRM THE FOLLOWING WITH THE F.B.I. THE SANDY SPRING'S DEVICE'S - GELATIN - DYNAMITE - POWER SOURCE 6 VOLT D BATTERY BOXES, DURACELL BRAND, CLOCK TIMER'S. THE MIDTOWN DEVICE'S ARE SIMILAR EXCEPT NO AMMO CAN'S TUPPERWARE CONTAINERS INSTEAD - POWER SOURCE SINGLE 6 VOLT LANTERN BATTERIES² DIFFERENT SHRAPNEL, REGULAR NAIL'S INSTEAD OF CUTT NAILS²

THE ABORTION CLINIC WAS THE TARGET OF THE FIRST DEVICE. THE MURDER OF 3.5 MILLION CHILDREN EVERY WILL NOT BE "TOLERATED"¹ THOSE WHO PARTICIPATE IN ANYWAY IN THE MURDER OF CHILDREN MAY BE TARGETED FOR ATTACK. THE ATTACK THEREFORE SERVES AS A WARNING ANYONE IN OR AROUND FACILITIES THAT MURDER CHILDREN MAY BECOME VICTIMS OF RETRIBUTION² THE NEXT FACILITY TARGETED MAY NOT BE EMPTY¹

THE SECOND DEVICE WAS AIMED AT AGENTS OF THE FEDERAL GOVERNMENT [E. A.T.F., F.B.I., MARSHALL'S E.T.C.] WE DECLARE AND WILL WAGE TOTAL WAR ON THE UNGODLY COMMUNIST REGIME IN NEW YORK AND YOUR LEGASLATIVE BUREAUCRATIC LACKEY'S IN WASHINGTON² IT IS YOU WHO ARE RESPONSIBLE AND PRESIDE OVER THE MUR OF CHILDREN AND ISSUE THE POLICY OF PEVERSION THAT DESTROYING OUR PEOPLE³ WE WILL TARGET ALL FACILITIES AND PERSONNEL OF THE FEDERAL GOVERNMENT² THE ATTACK IN MIDTOWN WAS AIMED AT THE SODOMITE BAR (THE OTHERSIDE). WE WILL TARGET SODOMITES, THERE ORGANIZATIONS, AND ALL THOSE WHO PUSH THEIR AGENDA³

IN THE FUTURE WHEN AN ATTACK IS MADE AGAINST TARGETS WHERE INNOCENT PEOPLE MAY BECOME THE PRIMARY CAUSALTIES A WARNING PHONE CALL WILL BE PLACED TO ONE OF THE NEWS BUREAU'S OR 911¹

Appendix 2

Text of case study: Lampley Hollow

Hello asshole! This is the eve of the bloodiest day in the history of Lampley Hollow!

You fucks want to step outside the law to show us much of a fuck your mother is? Well, you have attacked innocent people, and now innocent people will pay, on your behalf. And a few cops trying to stop us.

Sunday is the final day of Founders Day. On that day a minimum of 20 people will die there!

Here is how it will happen: Your department will receive a phone call ten minutes to the top of an hour, to announce the countdown. At the hour, the first explosion will occur.

Approximately six will die, mainly family members, and the bomber. This will start a panic, with people running in all directions. One of those directions will be toward the second bomber. Six seconds after the first explosion the second will occur, a distance from the first. Six more dead.

NOW for the big one! Two groups of people will collide, while escaping their respective explosions. At that time and place the third, largest explosion will occur. Eight dead, at least.

You wonder why we have people willing to do this and die over you? It's because they don't even know they are packing. And you cannot find them.

The people that die will even the score, and we start fresh. Don't fuckup or it will happen again. Perform your job with respect and dignity for the people you serve and you will save their lives. We regret this but feel an example of death is the only way to make you understand.

Substitutes "the bombings"

You remember the bomb in the planter last summer? That's right, the iron pipe bomb, with an electronic igniter. It was powered by four AA batteries in an Electronic supply pack, with a time delay. Don't count on a misfire this time. We worked out the ignition problems with that design.

It's a great day coming.

WESTERN CAPE

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

Appendix 3

Text of case study: Luke Jon Helder – pipe bombers

[Mailboxes are exploding!] [Why] [you] ask?²
[Attention people]¹
[You do things because you can] [and want (desire) to]²
[If the government controls what you want to do] [they] control what you can do!¹
[If you] are under the impression that death exists] [and you] fear it] [you] do anything to avoid
it?² [This is the same way pain operates.] [Naturally] [we] strive to avoid negative emotion/pain.]
[You] allow yourself to fear death!¹
[World authorities allowed] [and still allow] [you] to fear death!¹
[In avoiding death] [you] are forced to conform] [if you] fail to conform, [you] suffer mentally and
physically.³ [Are world powers] utilizing the natural survival instinct in a way that allows them
to capitalize on [the people]?¹

[To “live” (avoid death) in this society] [you] are forced to conform/slave away?²

[I’m here to help] [you] realize/understand that [you] will live no matter what!²
[It’s up to you] people to open [your] hearts and minds. [There is no such thing as death.] [The
people] I’ve dismissed from [this reality] are not at all dead!¹

[Conforming to the boundaries] [and restrictions] imposed [by the government] only reduces [the
substance] in your lives. [When 1% of the nation] controls 99% of [the nations] total wealth] [is it
a wonder] why there are control problems?²

[The United States] strives to provide freedom for [their] people. [Do we] really have personal
freedom? [I’ve] lived here for many years] [and I] see much limitation. [Does the definition of
freedom] include limitation? [I’ve] learned about the history of various civilizations in history]
[and I] see more and more limitation. [Do you] people enjoy [this trend of limitation]? [If not,
change it].¹

[As long as] [you] are uninformed about death] [you] will continue to say “how high”, when [the
government] tells [you] to “jump”.² [As long as] [the government] is uninformed about death] [they
will continue] tell [you] to “jump”. [Is the government] uninformed about death] [or are they
pretending]?²

[You] have been missing how things are.] [for very long].² [I’m] obtaining [your] attention in the only
way [I can].¹ [More info] is on its way! [More “attention getters”] are on the way! [If I could] [I
would change] only one person.] [unfortunately] [the resources] are not accessible. [It seems] killing
a single famous person would get the same media attention] [as killing] numerous un-famous
humans.] [There is] less risk of being abducted.] [Associated with] dismissing certain people.¹

[Sincerely,
Someone Who Cares]¹
[P.S. More info] will be delivered to various locations around [the country].¹

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP ATTRIBUTION IN L1/2 ENGLISH TEXTS

Appendix 4

Text of case study: Theodore Kaczynski (Unabomber)

The original text follows:

(Passage deleted at the request of the FBI)

This is a message from the terrorist group FC¹

We blew up Thomas Mosser last December because he was a Burston-Marsteller executive². Among other misdeeds, Burston-Marsteller helped Exxon clean up its public image after the Exxon Valdez incident. But we attacked Burston-Marsteller less for its specific misdeeds than on general principles.¹

Burston-Marsteller is about the biggest organization in the public relations field. This means that its business is the development of techniques for manipulating people's attitudes. It was for this more than for its actions in specific cases that we sent a bomb to an executive of this company.²

Some news reports have made the misleading statement that we have been attacking universities or scholars. We have nothing against universities or scholars as such.² All the university people whom we have attacked have been specialists in technical fields. (We consider certain areas of applied psychology, such as behavior modification, to be technical fields.)²

We would not want anyone to think that we have any desire to hurt professors who study archaeology, history, literature or harmless stuff like that. The people we are out to get are the scientists and engineers, especially in critical fields like computers and genetics. As for

the bomb planted in the Business School at the U. of Utah, that was a botched operation. We won't say how or why it was botched because we don't want to give the FBI any clues. No one was hurt by that bomb.¹

In our previous letter to you we called ourselves anarchists. Since "anarchist" is a vague word that has been applied to a variety of attitudes, further explanation is needed. We call ourselves anarchists because we would like, ideally, to break down all society into very small, completely autonomous units. Regrettably, we don't see any clear road to this goal, so we leave it to the indefinite future.²

Our more immediate goal, which we think may be attainable at some time during the next several decades, is the destruction of the worldwide industrial system.² Through our bombings we hope to promote social instability in industrial society, propagate anti-industrial ideas, and give encouragement to those who hate the industrial system.³

The FBI has tried to portray these bombings as the work of an isolated nut. We won't waste our time arguing about whether we are nuts, but we certainly are not isolated. For security

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

reasons we won't reveal the number of members of our group, but anyone who will read the anarchist and radical environmentalist journals will see that opposition to the industrial-technological system is widespread and growing.

Why do we announce our goals only now, through we made our first bomb some seventeen years ago? Our early bombs were too ineffectual to attract much public attention, or give encouragement to those who hate the system. We found by experience that gunpowder bombs, if small enough to be carried inconspicuously, were too feeble to do much damage, so we took a couple of years off to do some experimenting. We learned how to make pipe bombs that were powerful enough, and we used these in a couple of successful bombings as well as in some unsuccessful ones.

(Passage deleted at the request of the FBI)

Since we no longer have to confine the explosive in a pipe, we are now free of limitations on the size and shape of our bombs. We are pretty sure we know how to increase the power of our explosives and reduce the number of batteries needed to set them off. And, as we've just indicated, we think we now have more effective fragmentation material. So we expect to be able to pack deadly bombs into ever smaller, lighter and more harmless looking packages.

On the other hand, we believe we will be able to make bombs much bigger than any we've made before. With a briefcase-full or a suitcase-full of explosives, we should be able to blow out the walls of substantial buildings.

Clearly we are in a position to do a great deal of damage. And it doesn't appear that the FBI is going to catch us any time soon. The FBI is a joke.

The people who are pushing all this growth and progress garbage deserve to be severely punished. But our goal is less to punish them than to propagate ideas. Anyhow we are getting tired of making bombs. It's no fun having to spend all your evenings and weekends preparing dangerous mixtures, filing trigger mechanisms out of scraps of metal, or searching the sierras for a place isolated enough to test a bomb. So we offer a bargain.

We have a long article, between 29,000 and 37,000 words, that we want to have published. If you can get it published according to our requirements, we will permanently desist from terrorist activities. It must be published in the New York Times, Time or Newsweek, or in some other widely read, nationally distributed periodical.

Because of its length we suppose it will have to be serialized. Alternatively, it can be published as a small book, but the book must be well publicized and made available at a moderate price in bookstores nationwide and in at least some places abroad. Whoever agrees to publish the material will have exclusive rights to reproduce it for a period of six months, and will be welcome to any profits they may make from it.

After six months from the first appearance of the article or book it must become public property, so that anyone can reproduce or publish it. (If material is serialized, first instalment becomes public property six months after appearance of first instalment, second instalment, etc.)

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

[We] must have the right to publish in the New York Times, Time or Newsweek, each year for three years after the appearance of our article or book, three thousand words expanding or clarifying our material or rebutting criticisms of it.³

[The article] will not explicitly advocate violence. There will be an unavoidable implication that [we] favor violence to the extent that it may be necessary, since [we] advocate eliminating industrial society and [we] ourselves have been using violence to that end. But the article will not advocate violence explicitly, nor will it propose the overthrow of the United States Government, nor will it contain obscenity or anything else that [you] would be likely to regard as unacceptable for publication.³

[How do you] know that [we] will keep our promise to desist from terrorism if [our] conditions are met? It will be to our advantage to keep our promise. We want to win acceptance for certain ideas. If we break our promise, people will lose respect for us and so will be less likely to accept the ideas.²

[Our] offer to desist from terrorism is subject to three qualifications.¹

- First: [Our] promise to desist will not take effect until all parts of our article or book have appeared in print.¹
- Second: If [the authorities] should succeed in tracking us down and an attempt is made to arrest any of us, or even to question us in connection with the bombings, [we] reserve the right to use violence.³
- Third: [We] distinguish between terrorism and sabotage. By terrorism [we] mean actions motivated by a desire to influence the development of a society and intended to cause injury or death to human beings.² By sabotage [we] mean similarly motivated actions intended to destroy property without injuring human beings. The promise [we] offer is to desist from terrorism. [We] reserve the right to engage in sabotage.¹

[It may be just as well that] failure of our early bombs discouraged us from making any public statements at that time. [We] were very young then and [our] thinking was crude.² Over the years [we] have given as much attention to the development of our ideas as to the development of bombs, and [we] now have something serious to say.³ And [we] feel that just now the time is ripe for the presentation of anti-industrial ideas.²

[Please see to it that the answer to our] offer is well publicized in the media, so that [we] won't miss it. Be sure to tell us where and how our material will be published and how long it will take to appear in print, once [we] have sent in the manuscript. If the answer is satisfactory, [we] will finish typing the manuscript and send it to you. If the answer is unsatisfactory, [we] will start building our next bomb.²

[We] encourage [you] to print this letter.¹

FC

(Passage deleted at the request of the FBI)

Appendix 5

My work 7/89

[Why do I work?]

- [I use acquired knowledge]
- [I have also a degree of satisfaction from my work achievement]
- [necessity] - [as the salary which I earn allows me to live, travel] and [provides for my future when I stop work]

[Usually a working person spends three quarters of his useful life at work] [The person is involved with work problems for (10) most of his useful, productive life.] [After the person's education is completed or a certain level of knowledge is reached] [which allows to start work] till the retiring age, [all of life is dominated by work and its problems.] [For / majority of people this period last for thirty to forty years]

[During this time a person also builds his family, his life, outside interests:] [his whole reason for living, of course these can only be achieved if the work situation is normal, the income from work is steady and sufficient for all / expenses required] [to continue with one's life on a desired level.]

[In my opinion it is extremely important that the performed work gives the person satisfaction] [and enough pleasure to continue with it] [without too much stress].

[On the other hand the performed work must be

Page 2.

demanding enough] [to force the person to develop further]

[My work is mostly done in the same field, electricity] [with a number of changes during the years]

[I was and I am involve in electrical engineering heavy

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

current] and high voltage].²

[However, I was involved with some aspect of light current] as
[for a number of years I was involved in the electrical protec-
tion].² [Design for the protection for the substation rural or
urban]. [testing the designed protection scheme and comissioning].²
[My work gives me a degree of satisfaction] but also some
frustration].²

[I feel that there is a number of reasons for this:]

- [I am a woman who works in a man's world].¹
- [a certain degree of defficiency in my working style is (10)
not allowing me to perform the work excellently].¹
- [the poor comandments of English languages, which is not my
mother-tongue and Im not educated in it].²

[It is enough for an introduction]. [It is the time to say a
few words about my work as such].¹

[After receiving my degree as an electrical engineer]

Page 3.

[I completed my 'apprentice' training as an 'engineer in
training'] which is a practical training] [in the high voltage
laboratory of the transformer factory in Poland].³ (20)

[It was my first job] [and really highly challenging].² [The
transformers produced there were of a new design] [and they
required a full type test to be accepted by the client for
required purposes].²

[The laboratory was also brand new] [with special equipment
from which] [the most important was an impulse generator,
capable of producing 2,1 milion volts for a number of
mircoseconds]. (sic)³

[The team of people working there consisted of a number of
young engineers], [mostly friends] [as we completed our studies (in/...]

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

in the same technical university.²

Everybody there was very enthusiastic and dedicated and also our leader, our professor, knew how to deal with the young enthusiastic crowd of workers.²

The situation became different when I started work in South Africa, after receiving some extra training in England at English Electric in Stafford.²

When I look back, I can say honestly, I did not do too bad, as I was accepted by the fellow engineers, maybe as a kind of "odddity", and also by the bosses I worked for.³ (10)

My present work covers quite a broad field of the generator and auxiliaries and the generator system.²

Page 4.

Due to the different manufactures of the electrical generators which were installed over a number of years, in various power stations, I deal with quite a spectrum of different designed machines.²

Generally the generators differ in their basical size, generated power, cooling system, insulation material and in the regulation system.² The principle of generating the electrical power of course is the same for small and big generators.²

The generators installed in the older power stations have they MCR - (maximum continuous rating) of 30 MVA and in the new one 680 MVA and in the nuclear power station 920 MVA.³

The cooling systems of course are different, the cooling median used are: air, hydrogen and water.²

These require that the construction of the stator bars which carry generating current and situated in the stator iron slot are different.² As the cooling system must be more efficient (30)

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

for the higher current produced by the generator.]'

[Each manufacturer tries to introduce his own specific, patented, design.]'

[The aspects which make my work interesting are:

a) setting a specification for the generator which includes requirements for the future stations [end equipment installed in them].³

b) checking an answer [in a form of full proposal, to a tender issued and recommending [from the technical point of view, (10) the most successful tenderer, [and specifying why].³

[However the financial aspect of the tender is the taken into consideration.]'



UNIVERSITY of the
WESTERN CAPE

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

Appendix 6

TOPICS : STORY-JOB RELATED OR WORK RELATING STORY

[DURING MY EMPLOYMENT WITH A CONTRACTING FIRM CALLED "ENGINEER-
ING DRAFTING SERVICES"] [I WAS APPROACHED WITH MY AGENCY] [IF I
WOULD NOT LIKE TO CONSIDER AND FOR ESKOM] [BECAUSE THEY HAVE BEEN
ADVERTISING A VACANT POST IN THE LIGHTING SECTION]⁴
[THOSE TIME I HAVE HAD A SHORT TIME-TERM CONTRACT WITH A CONSULT-
ING ENGINEERING] [WHICH WAS LEADING TO THE END]² [WHILE I WAS
EMPLOYED WITH THE CONSULTING ENGINEERS] [I HAVE BEEN INVOLVED (10)
ALSO IN THE LIGHTING FIELD FOR DOMESTIC INSTALLATION] [SUCH AS:
SHOPPING CENTRE, DUPLEXES, SCHOOLS, HOTELS, PRIVATE RESIDENCES]².
[WHEN I HEARD THAT I COULD HAVE A LONG TERM CONTRACT WITH A SUCH
A BIG COMPANY AS ESKOM I WAS A LITTLE FRIGHTENED] [BUT I HAD
EXCEPTED ALSO] [BECAUSE I KNEW FEW CZECHOSLOVAKIAN PEOPLE IN
THERE]³ [IT ALL HAPENED EXACTLY IN FEBRUARY 1978] [WHEN I STARTED
WORKING FOR ESKOM IN MEGAWATT PARK, SITUATED IN SANDTON-RIVONIA]²
[THE COMPLEX WAS SO BEATIFUL] [IT HAD FASCINATED EVERY NEWCOMMER]²
[THE WORKSTATION WAS ALL OPEN PLAN OFFICES] [WITH VERY MODERN
FURNITURES AND MOST BEATIFUL PLANTS I HAVE EVER SEEN IN MY (20)
LIFE]² [THE SECTION WHICH I HAVE JOIN WAS FAIRLY BIG] [AND IT WAS
RESPONSIBLE FOR THE DESIGN OF THE WHOLE POWER STATION LIGHTING]
Page 2.
[WHICH CONSISTED OF 1) INTERIOR LIGHTING OF BUILDINGS SUCH AS
OFFICES, CONTROL ROOMS, LECTURE ROOMS, DINNING AREAS, LABORA-
TORIES, RECEPTION AREAS, WORKSHOPS AND WILD VARIETY OF OTHER
BUILDING LIGHTING]³ [WITH THIS KIND OF WORK I WAS QUITE FAMILIAR]
[AND I FELT CONFIDENT AND COMPLETED MY WORK WELL TO MY SUPER-
VISOR SATISFACTION]² [BUT WHEN IT COME TO DO THE LIGHTING
DESIGN OF THE OUTSIDE PLANTS] [SUCH AS TURBINE HOUSE, BOILERS
HOUSE, PRECIPITATORS, COAL AND ASH DRIVE HOUSES AND TRANSFER (30)

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

HOUSES AND THEIR RELEVANT CONVEYORS RUNS¹ [THOSE ASSIGNMENTS WERE VERY MUCH MORE COMPLICATED] AND REQUIRED ALOT OF ENGINEERING INPUT² [THE LIGHTING DESIGN OF THOSE PLANTS INVOLVED REGULAR SITE VISITS, CLEAR AND UNDERSTANDABLE COMMUNICATION WITH ALL RELEVANT DISCIPLINES INVOLVED IN THE PROJECT] [SUCH AS ARCHITECTURAL, MECHANICAL, CIVIL A/C-ING, TELECOMMUNICATION AND OF COURSE THE CONTRACTORS DISCIPLINES.³ [IT ALSO REQUIRED of THE STUDY / HOW THE PLANTS FUNCTIONING, STUDY OF THE CONSTRUCTION DRAWINGS, ABILITY TO READ THE MAKERS DRAWING] (10) [TO BE ABLE TO PRIPARE A GENERAL ARRANGEMENT DRAWINGS OF THE PLANTS] [AND THEN THE REQUIRED LIGHTING FOR GOOD VISIBILITY] [GOOD SECURITY AND SAFETY] [AND ALSO BEAUTY]

Page 3.

[AND EMPHASIS OF THE PLANT]⁶ [THE PROPER LIGHTING DESIGN COULD ONLY BE ACHIEVED WITH A SYSTEMATIC APPROACH, PROPER SITE MEASUREMENTS AND SITE MARK-UPS] [AND OBTAINING THE ALL UP-TO DATE INFORMATIONS.³ [THIS TYPE OF WORK WAS VERY INTERESTING] [CHALENGING] [AND VERY PROMISING.³ [THE TYPE OF WORK], [WHICH I WAS INVOLVED IN] [AND THE LOVELY PEOPLE I WORKED WITH] [AND A (20) PLEASANT ENVIRONMENT AND ATMOSPHERE] [MADE ME TO CHANGE MY MIND] [AND BECOME A PERMANENT EMPLOYEE TO ESKOM.⁵ [THEN IN OCTOBER 1978 I STARTED ON THE PERMANENT BASIS] [AS A SENIOR DRAUGHTS-WOMAN] [AND ALL THE PEOPLE EXCEPTED ME WITH WARM WELCOME TO BECOME ONE OF THEM.³ [AS THE TIME PASSED BY], [I HAVE GAINED MORE EXPERIENCE] [AND ALSO GOT INVOLVED IN MORE DIFFICULT ASSIGNMENTS].² [AFTER MANY OF SUCSESFUL JOB COMPLETIONS ON THE COAL POWER STATION] [I RECEIVED A COMPLIMENTS] [WHICH MADE ME EVEN MORE ENTHUSIASTIC] [AND MORE INTERESTED IN MY WORK]³ [AFTER FEW YEARS OF WORKING ON THE COAL POWER STATION] [I HAVE BEEN ASKED FROM (30)

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

MY DIVISION MANAGER IF I WOULD NOT TO LIKE TO GET INVOLVED IN
THE LIGHTING DESIGN FOR A HYDRO STATION.³ [I DID EXCEPT THE
NEW CHALLENGE] AND STARTED WORKING ON THE DRAKENSBERG PUMPED
STORAGE SCHEME.²

Page 4.

[I MUST SAY IT WAS QUITE AN ADVENTURE] AND FAN IS WELL, [WHEN WE
VISITED THE SITE WE HAD TO CHANGE TO THE RUBBER CLOTHING RUBBER
BOOTS] [BECAUSE THE TUNNELS THROUGH WHICH WE HAD TO ACCESS THE
UNDERGROUND ENVIRONMENT WAS ALWAYS FULL OF WATER, DURING THE
EXCAVATION.]⁵ [WE REALY LOOKED LIKE PEOPLE FROM THE MOTHER (10)
PLANET] [MEANWHILE WE WERE WORKING UNDER THE GROUND COUPLE OF
HUNDREDS METERS]² [I WAS INVOLVED IN THIS HYDRO PROJECT RIGHT
TO THE END] [AND ALSO IT BECAME ONE OF MY BIGGEST SUCCESSFUL
ACHIEVEMENTS]² [AFTER THE COMPLETION OF THIS PROJECT] [I HAVE
BEEN PROMOTED TO THE DESIGN DRAUGHTSWOMAN OFFICIALLY] [EVEN
THE RESPONSIBLE DUTY I HAVE BEEN CARRYING OUT LONG TIME
ALREADY]² [IT WAS A VERY NICE AND PLEASANT FEELING] [AND I WAS
VERY PROUD OF IT.]² [THE WORK SITUATION HAD CHANGED A BID] [BUT
MAJORITY WAS STILL THE SAME]² [I WAS MORE INVOLVED FOR
PREPARATION WORK FOR THE WHOLE SECTION] [MORE DESIGN (20)
ORIENTATED FOR THE INDIVIDUALS] [AND ALSO WAS INVOLVED IN ESKOM
GENERAL DATA FOR STANDARDS]³ [MY WORK BECAME MORE ADMINISTRA-
TIVE]¹ [THEN IN 1986 I HAVE ATTENDED AN ILLUMINATION ENGINEER-
ING COURSE] [HELD AT TECHNIKOM AND OR-

Page 5.

GANISED BY "ILESA"] [WHICH I COMPLETED SUCCESFULLY]³ [I MUST
ADMIT IT, THAT ONLY AFTER I COMPLETED THIS COURSE] [I HAVE FOUND
OUT HOW MUCH MORE THE WORD OF LIGHTING REALY MEANT] [HOW MANY
SECRETS LIE IN THE DESIGN] [AND HOW LITTLE I REALY KNEW]⁴ [I MUST
SAY THAT IT CHANGED MY OWN APPROACH TO THE DESIGN PARAMETERS (30)
FOR /

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

FOR THE LIGHTING.² [DURING THE YEAR OF THE COURSE WE LEARNED]
[THE PHYSIOS OF THE HUMAN EYE, THE SPECTRUM, PHOTOMETRY, COLOUR
REENDERING, LAMP AND LUMEN STRUCTURES, STREET LIGHTING, ROAD
LIGHTING, CHURCHES, MUSLUM AND GALLERIES, SPORT FLOODLIGHTIIG
DAYLIGHTING, SUPPLEMENT ARTIFICIAL LIGHTING] AND MANY MANY MORE
INTERESTING AND FASCINATING FACTORS NEEDED TO A PROPER LIGHT-
ING DESIGN]³ [AFTER I COMPLETED THIS COURSE] [I HAVE TRIED TO
TRANSFER OR MY NEW EXPERIENCE ON TO MY COLLEGES].¹ [IT WAS
DIFFICULT IN THE BEGINING] [BUT SLOWLY BECOME RECOGNISBLE].²
[THEN A YEAR LATER] [IN MARCH 1987] [I WAS PROMOTED TO THE SECTION (10)
LEADER OF THE LIGHTING SECTION] [WHICH I HELD UP TO NOW] [AND MY
JOB MISSION IS: TO PROVIDE TECHNICAL SERVICES IN THE FIELD OF
THE LIGHTING] [AND SMALL POWER FOR VARIOUS POWER STATIONS: MAJUBA,
KENDAL, MATIMBA].⁵

UNIVERSITY of the
WESTERN CAPE

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

Appendix 7

Gentlemen

[I must admit that] [I read your newspaper rather occasionally] [and when I do so I am rather disappointed].² [All the articles on the first page describe sensational events] [which are rather of a mean or no value to me] [and to other readers as well, I hope].³ [Is it done on purpose?]¹ [Is this a way to attract attention to something less important] [and make the readers forget about some real problems in our country?]² (10)

[Have you been on Church street] [during lunch time recently?]¹ [Have you seen all the baggers?]¹ [What have you done to help them] [to change their lives?]²

[Should we not talk about special schools for black disabled people?]¹ [Even they can be useful in a society] [but we must create an opportunity for them to learn them a trade].² [It does not to be something complicated] [but simple] [and useful].² [It will improve their quality of life] [and ours at the same time].²

[With regards] UNIVERSITY of the WESTERN CAPE (20)

Page 2.

[On the day of my retirement] [I will pack my suitcases] [and go to my country side house].²

[My country side house will be an old farm house] [with a garden full of flowers].² [All the goods inside my house will be made of a wood] [and other natural materials]. [There will not be place for any plastic!]¹

[During my retirement] [I would like to do all the things for which I have not had time before].¹ [Books reading, pottery, drawing] [perhaps even painting]. [But I would like to have somebody / (30)

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

somebody to talk with, [to exchange opinions] and to argue.²
To do that will be a great pleasure but I would like to be
somhow useful to other people as well.² [How?] [I do not know
yet] but I have still some time to think about it.²

Page 3.

Pollution - this is one of the biggest problems of our
century and one should think what price we are going to pay
for it.²
[But even bigger problem is that] [so many people are not aware
of the consequences of the pollution].² [Many others are aware] (10)
[but money are more important for them].² [Unfortunately in SA
does not exist any law] which will force the industry to take
some precaution measures.¹

Page 4.

4. Conservation

[Every generation should think about other generations] which
will come after and which will have to live in the world the
previous generation have left.² [We should use the nature to
our addvantage] but not damage it.² [We should let our children
and gradchildren] enjoy it as well.² [Use does not mean abuse] (20)
[therefore everything in the nature should be used sparingly]
[and a great care should be taken to restore that what was used].³
[Now days we are able to predict the consequences of our
activities] and more attention should be paid to plan a proper
method [to deminish the negative sides of existence].³ [It sounds
very pompes].¹

Page 5.

[I am afraid that] [I can not think of anything what will keep
the prices of food down].² Should the government subsidized the
food? [the high prices] of food production are connected to (30)
the/...

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

the high prices of anything else [e.g. transport costs, labour at cetera]. [Are the big supermarkets a good solution for a food distribution?] [Will small shops not be better?] [The running cost of a supermarket must be very high] [and a cost of wasted (too old for sale) food must be included in the price].² [Small shops are normally more elastic in a decision making process] [better organized] [and less food is wasted in them].² [A good developed network of small shops should create more competition] [better quality] [and lower prices].¹

Page 6.

(10)

[The government should do something to make people to lead their lives on a level they can afford]. [The very high inflation is between others] [due to big credits people live on] [due to spending money which do not exists].² [But this does not concern only people in the country] [but the country itself].¹ [What did happen to SA economy?] [Do we export enough goods] [to import everything we need].²

[Another problem is the society awarness] [and understanding of economical problems].² [This is a big role for TV, radio and press to play].¹

(20)

Page 7.

[I start every day at 5h45] [with phisical eccesizes]. [I try do make myself tired] [and it takes me about 15 minutes].² [Then, according to my son,] [I run around doing nothing a half asleep a half awaked].² [But still during this time] [I prepare a breakfast for my son and me] [dress myself] [and at 7h00 I am ready to leave for my work].³ [I start work at 7h45] [with checking my direy] [where I write all appointments] [and things to be done on that day].² [I make a lot of effort] [to work until lunch time].¹ [But not every day I have time for lunch time] [and then I work (30

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

through untlil afternoon teetime]²
[My work ends normally at 16h45]¹ [After work I rush home] [then
prepare food for my son] [who is always very hungry]² [At home]
[I start the second shift of my work]¹ [I play tennis on
Mondays, Thusdays] [and sometimes Fridays]¹



UNIVERSITY *of the*
WESTERN CAPE

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

Appendix 8

Dear Sirs,

[It remains an undeniable fact] that there are millions of hungry people in our country.² [It is the result of commonly spread inconsistency of the white race] to help to natives [by means of food supply, medical care], to make them grow in number [i.e. to destroy the natural rate of population growth], and then to try to help the hungry artificially multiplied population.⁴ [Of course we are far from admittance to the guilt:] [oblivious to the real reason] [which has produced millions of the poor] and hungry [we try to do our hypocritical best to help them] by throwing odds and ends from our fully packed fridges.³ [But still as long as we do that out of guilty conscience] [we are being punished enough].²

[The only solution to the problem] [which comes to my mind] [is as surrealistic as the cause of the problem:] [why not to supply the hungry with invisible caps] [so they could come sometimes to our homes] [eat to their heart contents without arising our righteous indignancy to such a outrageous behaviour].⁴

Page 2.

[The day when I retire appears to be to most glamorous] [the most dreamt of day of my life].² [I think I have so many expectations,] [planned activities,] [a general feeling of long awaited freedom] [to do whatever I please] [that it all makes difficult to define something in detail].⁴ [which is wise since I shall probably change] [and so will my interests].² [Had the day be today I would not have been writing this paper for Hilton] (for sure) [but I would be bursting with joy tossing in/... (30)

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

in my garden, fondling flowers, harring up the buds, [trying to communicate with grass and trees and clouds, guessing their wishes] and needs [and trying to satisfy them].⁴ [But of course things may happen: [I may fall ill (what about quite common at retiring age Alzheimer disease?) [or in any other way disabled] and then it is better to be right where I am today].⁴

Page 3.

[In general] [I think that pollution is the thing which should not have happened at all] [(which means that perhaps the three last centuries should not have happened)].² (10)

[I have been impressed by learning the theory about GAJA (the great goddess of Earth) [which assumes that earth is a living organism].² [Although the idea seems to be shocking (to say the least)] [the scientists who favore it] were able to find so many indisputable proves to support it] (the same constant temperature around the earth from billions of years for instance) [that no reasonable human being does not reject it] any longer.³ (20)

[When I think about pollution I have to ask the following question on behave of Gaje: [is there a safe place from mankind?]²

Page 4.

[Conservation should be treated as the most vital] of our activities [if we wish to survive as a spicy].² [We should become more humble towards nature].¹ [We should restrain at once our needs] [which seems to be blown up beyond every reasonable limit] [meeting our hyper needs we tend to produce more at the expence of more and more robbed and violated nature].³ [We should/... (30)

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

should undertake any industrial] (i.e. potentially disastrous
to nature) [activity only when we are able to nihilate the
harmful to our environment effects].¹ [Then] [and only then] [we
might be able to think that the future generations will have
fresh air to breath, nutritshious helthy food to eat and
clear water to drink] [(to say nothing about marvellous natural
surroundings to live in)].²

Page 6.

[The first solution which immediately cross one's mind is]: (10)
[to take from the reach and to give to the poor]; [but this
would be a very shortlasting remeady]² [And of course in the
end] [there would be the newly reach and the newly poor] [so the
whole prosses should have to repeated ceaselessly] [which seems
to be boring and tiresome].³ [What perhaps could be a solution
is the existance of a honest, dedicated and wise government]
[with a sound economic policy].² [Do you know of one?]
[The high prices would eventually affected the whole population]
[although it seems obvious that those who are interested in
(buying a Sunflowers by Van Gogh at R50 000 000] would (20)
(suffer slightly less than the rest of society)].²

Page 5.

[Here it is: let's have an agreement with the price levels] [(or
let's force them) not to change the digits they bear with time].²
[Let's make the lebelns indifferent to any,] [even the most
considerate or on the contrary: most cruel,] [measures undertaken
in order to change what they have read from their birth].³ [And
they could even learn how to cheat the oppressors.] [at first
they would pretend to have the new numbers printed on them] [but
the moment the price putting men goes away the lebelns would (30)
assume/...

INVESTIGATING A PROPOSED STANDARDISED METHOD FOR AUTHORSHIP
ATTRIBUTION IN L1/2 ENGLISH TEXTS

assume the previous price.³ [Our future prosperity lies in the
ingenuity of price labels.]¹

(A) from the previous page, pse.

Page 7.

[My weekly schedule] seems to be unbearably boring. [The days
are so enjoyably similar to each other.]¹ [But if you see what
are (they) filled up with] [you will understand why I have said
"enjoyably similar".]² [Evry working day I get up at six and
immediately check on the weather] [which is silly] [because I (10)
know I will enjoy any kind of the sky mood.]³ [Then if it is not
raining] [I go out and run for a couple of minutes in (the garden)
[greeting everything what I can see on my way.]³ [Then I get
ready to go to work] [which is very creative since I have nothing
to put on] [but still I have to dress.]³

[Every day I take different way to work] [I mean phisically
it's the same streets] [but not by what I see.]³ [Every day and
every moment of my driving ups and downs] [I see different hills
and woods, roofs, windows and magnolias, different horizons
and absolutely amazingly different colours of everything.]² (20)
[Then I work with greater or lesser pleasure which depends on
my mood] [and also on (the subject) I am busy with.]²