

**UNIVERSITY OF KWAZULU-NATAL**

**FORENSIC COMPUTING STRATEGIES FOR ETHICAL ACADEMIC WRITING**

**By  
Sashen Govender  
204002155**

**A dissertation submitted in fulfilment of the requirements for the  
degree of  
Master of Commerce in Information Systems & Technology**

**School of Information Systems & Technology  
Faculty of Management Studies**

**Supervisor: Prof. Rembrandt Klopper**

**2009**

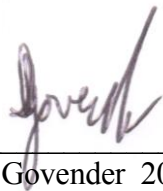
## **DEDICATION**

*I would like to dedicate this to my parents Logie and Dinky, my sister Nishantha, my Uncle Peggs and Aunty Menie and finally, Suheena, for their undying support and assistance throughout the duration of this project.*

## DECLARATION

I .....**Sashen Govender** .....declare that

- (i) The research reported in this dissertation/thesis, except where otherwise indicated, is my original research.
- (ii) This dissertation/thesis has not been submitted for any degree or examination at any other university.
- (iii) This dissertation/thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
- (iv) This dissertation/thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
  - a) their words have been re-written but the general information attributed to them has been referenced:
  - b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced.
- (v) This dissertation/thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation/thesis and in the References sections.



---

S. Govender 204002155

20 October 2009  
Date

## ACKNOWLEDGEMENTS

Having completed my research, I would like to express my sincere appreciation to Professor *Rembrandt Klopper*, my supervisor, for his expertise, time and meticulous attention to detail. I cannot thank you enough.

## ABSTRACT

University of KwaZulu-Natal

READING BETWEEN THE LINES:

FORENSIC COMPUTING STRATEGIES FOR ETHICAL ACADEMIC WRITING

Sashen Govender

This study resulted in the creation of a conceptual framework for ethical academic writing that can be applied to cases of authorship identification. The framework is the culmination of research into various other forensic frameworks and aspects related to cyber forensics, in order to ensure maximum effectiveness of this newly developed methodology. The research shows how synergies between forensic linguistics and electronic forensics (computer forensics) create the conceptual space for a new, interdisciplinary, *cyber forensic linguistics*, along with forensic auditing procedures and tools for authorship identification. The research also shows that an individual's unique word pattern usage can be used to determine document authorship, and that in other instances, authorship can be attributed with a significant degree of probability using the identified process. The importance of this fact cannot be understated, because accusations of plagiarism have to be based on facts that will withstand cross examination in a court of law. Therefore, forensic auditing procedures are required when attributing authorship in cases of suspected plagiarism, which is regarded as one of the most serious problems facing any academic institution.

This study identifies and characterises various forms of plagiarism as well the responses that can be implemented to prevent and deter it. A number of online and offline tools for the detection and prevention of plagiarism are identified, over and above the more commonly used popular tools that, in the author's view, are overrated because they are based on mechanistic identification of word similarities in source and target texts, rather than on proper grammatical and semantic principles.

Linguistic analysis is a field not well understood and often underestimated. Yet it is a critical field of inquiry in determining specific cases of authorship. The research identifies the various methods of linguistic analysis that could be applied to help

establish authorship identity, as well as how they can be applied within a forensic environment. Various software tools that could be used to identify and analyse source documents that were plagiarised are identified and briefly characterised. Concordance, function word analysis and other methods of corpus analysis are explained, along with some of their related software packages. Corpus analysis that in the past would have taken months to perform manually, could now only take a matter of hours using the correct programs, given the availability of computerised analysis tools.

This research integrates the strengths of these tools within a structurally sound forensic auditing framework, the result of which is a conceptual framework that encompasses all the pertinent factors and ensures admissibility in a court of law by adhering to strict rules and features that are characteristic of the legal requirements for a forensic investigation.

## TABLE OF CONTENTS

Chapter One	
Problem Statement And Research Design	
1.1 Introduction	16
1.2 The Term “Conceptual Framework”	16
1.3 The Term “Ethical Writing”	17
1.4 The Relationship between Ethical Writing and Cyber Forensics	17
1.5 Motivations for the Study	18
1.6 Problem Statement	19
1.6.1 Subproblem 1	22
1.6.2 Subproblem 2	22
1.7 Aim and Objectives	22
1.7.1 Objective 1	22
1.7.2 Objective 2	23
1.8 Interim Research Question and Subquestions	23
1.8.1 General Research Question	23
1.8.2 Subquestion 1	23
1.8.3 Subquestion 2	23
1.9 Envisaged Contribution to the Body of Knowledge	23
1.10 Research Design	24
1.11 Layout of the Study	24
1.12 Conclusion	24
Chapter Two	
Literature Review	
2.1 Introduction	26
2.2 Cyber Forensics and the Nature of Electronic Material	26
2.3 The Role of Language Analysis in Forensic Auditing	28
2.3.1 Using Syntactic Patterns in Forensic Analysis	29
2.3.2 Using Word Patterns in Forensic Analysis	32
2.3.3 Word Frequency Analysis	42
2.4 Ethical Writing	42
2.5 Referencing	43
2.6 Plagiarism	44
2.6.1 Common Knowledge	45
2.6.2 Types of Plagiarism	45
2.6.2.1 Idea Plagiarism	46
2.6.2.2 Text Plagiarism	46
2.6.2.3 Patch Writing and Paraphrasing	49
2.6.3 Collusion	51
2.6.4 Self-Plagiarism	52
2.6.4.1 Redundant and Duplicate Publications	52
2.6.4.2 Academic Self-Plagiarism	53
2.6.4.3 Salami Slicing (Data Fragmentation and Augmentation)	53
2.6.5 Effects of Plagiarism at Academic Institutions	53
2.6.5.1 Plagiarism causes Grade Inflation	53
2.6.5.2 Plagiarism hinders the Proper Pitching of Assignments	54
2.6.5.3 Plagiarism Cases at academic institutions are Huge Time Sinks	54

2.6.5.4 Encountering Plagiarism is depressing for Academic Institutions	54
2.6.5.5 Plagiarisers undermine their own Education	54
2.6.5.6 Successful plagiarism encourages Lifelong Dishonesty	54
2.6.5.7 Rampant plagiarism demoralises Honest Students	54
2.6.6 Techniques in the Deterrence of Plagiarism	55
2.6.6.1 Add Serious Anti-Plagiarism Warnings to the Syllabus	55
2.6.6.2 Tell Students how the institution detects Plagiarism	55
2.6.6.3 Make it known that the university uses Plagiarism-Detection Software	55
2.6.6.4 Inform students plagiarisers will have to appear before a Judicial Committee	55
2.6.7 Techniques to Detect Plagiarism	56
2.6.7.1 Software	56
2.6.7.1.1 CopyCatch Investigator	57
2.6.7.1.2 SafeAssign (MyDropBox)	58
2.6.7.1.3 Eve2	59
2.6.7.1.4 Glatt Plagiarism Screening Program	59
2.6.7.1.5 JPlag	60
2.6.7.1.6 PlagiarismDetect	60
2.6.7.1.7 Pl@giarism	61
2.6.7.2 Outsourcing and Online Solutions	62
2.6.7.2.1 <i>Google</i> as an Anti-Plagiarism Tool	64
2.6.7.2.2 CopyScape	65
2.6.7.2.3 Article Checker	65
2.6.7.2.4 Dupli Checker	65
2.6.8 Prosecution of Offenders	65
2.6.9 Plagiarism and Cultural Relationships	66
2.7 Stylistics and Linguistics	66
2.7.1 Corpus-Based Text Analysis	67
2.7.2 Corpus Annotation	68
2.7.3 Function Word Analysis	69
2.8 Models and Tools used in e-Forensics	72
2.9 Legal Implications	72
2.10 The e-Forensic Life Cycle	72
2.10.1 Understanding the Cyber Forensic Investigation Process	73
2.10.2 MD5 and CRC Authentication	74
2.11 Forensic Frameworks	74
2.11.1 The Processes and Factors involved in a Digital Forensic Investigation	75
2.11.2 Elements when Planning a Computer Forensic Investigation	76
2.11.2.1 Data Protection	76
2.11.2.2 Data Acquisition	76
2.11.2.3 Imaging	76
2.11.2.4 Data Extraction	77
2.11.2.5 Data Standardisation for use by Non-Technical Persons	77
2.11.2.6 Reporting on the Findings	77
2.12 Crime Scene Investigation Framework	77
2.13 Digital Crime Scene Investigation Framework	79
2.13.1 The System Preservation and Documentation Phase	79
2.13.2 The Evidence Searching and Documentation Phase	80



2.13.3 Digital Event Reconstruction	82
2.14 Cyber Tools online search for Evidence (CTOSE)	83
2.15 Digital Forensic Framework that incorporates Legal Issues (FORZA)	84
2.15.1 Fundamental Principles in Digital Forensics Investigation Procedures	84
2.15.2 The Participants in the FORZA Framework	86
2.15.3 The Six Key Questions	87
2.15.4 Implementing the Legal Aspects	89
2.16 A Conceptual Framework for the Forensic Auditing of Academic Assignments	91
2.17 Content Analysis	92
2.17.1 Sampling Units	93
2.17.2 Context Units	93
2.17.3 Recording Units	93
2.18 Character Recognition and Plagiarism	93
2.18.1 Form Identification	94
2.18.2 Field Isolation	94
2.18.3 Segmentation	94
2.18.4 Recombining Segments	94
2.18.5 Recognition	94
2.18.6 Organising Character Candidates	94
2.18.7 Dictionary Based Correction	94
2.18.8 Levels of Acceptance	95
2.18.9 Rejection of the Result	95
2.18.10 Summary	95
2.19 Anti-Plagiarism Tools	96
2.19.1 Online Tools	96
2.19.2 Offline Tools	96
2.20 Conclusion	96
Chapter Three	
Forensic Linguistics	
3.1 Introduction	98
3.2 Linguistic Analysis	102
3.2.1 Auditory Phonetics	103
3.2.2 Acoustic Phonetics	103
3.2.3 Semantics	104
3.2.4 Discourse Analysis	104
3.2.5 Pragmatics	104
3.2.6 Stylistics	105
3.2.7 Language of the Law	106
3.2.8 Language of the Courtroom	106
3.2.9 Interpretation	106
3.3 Linguistic Variation	108
3.3.1 Language Variance as markers of Social Class or of Individual Language Use	109
3.3.1.1 Associations	109
3.3.1.2 Identification of Class Features	109
3.3.1.3 Individuation	109
3.3.2 The terms “Language” and “Dialect”	110
3.3.3 The term “Idiolect”	110
3.4 Linguistic Variation Analysis Techniques	111

3.4.1 Linguistic Variables	111
3.4.2 Sentence Variation	112
3.5 Relationships between Forensic Linguistics and Other Elements	113
3.5.1 Relationship between DNA and Language Stylistics	113
3.5.2 Relationship between Linguistics and Document Examination	114
3.5.3 Relationship between Linguistics and Software Forensics	114
3.6 Stylistics and Forensics	114
3.6.1 Linguistic Stylistics	115
3.6.2 Analysis Techniques	116
3.6.2.1 Software Tools for establishing Concordance	117
3.6.2.1.1 Phrase Context	117
3.6.2.1.2 Concordance	118
3.6.2.1.2 MonoConc Pro and Paraconc	119
3.6.2.2 Software Tools for Word Frequency Analysis within Corpora	120
3.7 Framework for Authorship Identification Analysis	134
3.7.1 Organisation of the Case	135
3.7.2 Understanding the Problem	135
3.7.3 Method of Investigation	135
3.7.4 The Format for a Forensic Audit Report	136
3.7.5 Kings College London Approach to Text Analysis	137
3.8 Summary	139
Chapter Four	
A Conceptual Framework For Ethical Academic Writing	
4.1 Introduction	140
4.2 Cyber Forensic Linguistics	141
4.3 General Conceptual Framework for Forensic Linguistic Analysis	142
4.4 Similarities between the General Framework for Problem-Solution Oriented Research and the General Conceptual Framework for Forensic Linguistic Analysis	146
4.5 The Framework Applied to an Hypothetical Scenario	148
4.5.1 Initiating Step 1	148
4.5.2 Initiating Step 2	149
4.5.3 Initiating Step 3	149
4.5.4 Initiating Step 4	149
4.5.5 Initiating Step 5	149
4.5.6 Core Step 1 - Identify the problem	149
4.5.7 Core Step 2 – Gather Data	150
4.5.8 Core Step 3 – Analyse the Data	150
4.5.9 Core Step 4 – Produce Results	151
4.5.10 Core Step 5 – Create Report	151
4.6 Summary	151
Chapter Five	
Conclusions And Recommendations	
5.1 Introduction	153
5.2 Recap/overview of the Research	153
5.2.1 Chapter 1 - Introduction	153
5.2.2 Chapter 2 – Literature Review	153
5.2.3 Chapter 3 – Forensic Linguistics	154
5.2.4 Chapter 4 – A Conceptual Framework for Forensic Linguistic Analysis	154

5.3 Research Questions Revisited	155
5.3.1 Introduction	155
5.3.2 General Research Question	155
5.3.3 Subquestion 1	155
5.3.4 Subquestion 2	157
5.4 The Importance of Forensic Linguistics in Cyber Forensic Analysis	157
5.5 Conclusions and Recommendations	158
5.5.1 Conclusions	158
5.5.2 Recommendations	159
BIBLIOGRAPHY	161

## TABLES AND FIGURES

<i>Number</i>	<i>Page</i>
Figure 1.1: Framework for problem-solution oriented research, from Klopper (2008) (Contents SIC) .....	21
Figure 2.1: The general pattern of English transitive sentences .....	29
Figure 2.2: A typical English noun phrase consisting of a predetermining title and a surname as the core constituent of the phrase.....	30
Figure 2.3: A typical English noun phrase consisting of a personal pronoun as core constituent and a numeral as post-determining constituent. ....	30
Figure 2.4: An English noun phrase consisting of a predetermining indefinite article, an adjective indicating size, an adjectival phrase consisting of an adjective indicating degree, and an adjective indicating colour, as predetermining constituents, with a proper name as the core constituent of the phrase. ....	31
Figure 2.5: An English prepositional phrase consisting of a preposition as the core constituent of the phrase, and a noun phrase indicating an instrument as post-determining constituent.....	32
Figure 2.6: An English prepositional phrase consisting of a preposition as the core constituent of the phrase, and a dependent clause as post-determining constituent. ...	32
Figure 2.7: Colour coding for main lexical categories, adopted from: visualthesaurus.com (2009) .....	34
Figure 2.8: The adjectival antonyms “evil” and “good”, adopted from: visualthesaurus.com (2009) .....	34
Figure 2.9: “Pen”, adopted from: visualthesaurus.com (2009).....	34
Figure 2.10: “Audit”, adopted from: visualthesaurus.com (2009).....	35
Figure 2.11: “Analysis”, adopted from: visualthesaurus.com (2009).....	36
Figure 2.12: “Electronic Computer”, adopted from: visualthesaurus.com (2008) .....	37
Figure 2.13: “Cyberbate”, adopted from: visualthesaurus.com (2008) .....	37
Figure 2.14: “Writing”, adopted from: visualthesaurus.com (2008) .....	38
Figure 2.15: “Ethical Motive”, adopted from: visualthesaurus.com (2008).....	38
Figure 2.16: “Ethical”, adopted from: visualthesaurus.com (2008) .....	39
Figure 2.17: “Crime”, adopted from: visualthesaurus.com (2008).....	40
Figure 2.18: “Cybercrime”, adopted from: visualthesaurus.com (2008).....	40
Figure 2.19: “Fraud”, adopted from: visualthesaurus.com (2008) .....	41
Figure 2.20: Turnitin’s Plagiarism Prevention System, adopted from: Turnitin.com (2004).....	63
Figure 2.21: Caval project process, from: O’Connor - Cheating and electronic plagiarism [2003] .....	64
Table 2.1: Function Word Usage in the Oxford English Corpus, from: AskOxford.com (2007a) .....	70
Table 2.2: The 100 most common Function Words in Oxford English Corpus, from: AskOxford.com (2007a) .....	70
Table 2.3: Top 25 Function Words Classified as Nouns, Verbs and Adjectives in Oxford English Corpus, from: AskOxford.com (2007a) .....	71
Figure 2.22: Major Categories of Phases in the Crime Scene Investigation Framework, adopted from: Carrier and Spafford (2004).....	78
Figure 2.23: The Digital Crime Scene Investigation Phases, from: Carrier and Spafford (2004).....	79

Figure 2.24: The Evidence Searching Phases, from: Carrier and Spafford (2004) .....	81
Figure 2.25: Evidence Reconstruction Phase, from: Carrier and Spafford (2004) .....	82
Figure 2.26: CTOSE Reference Process Model, from: Broucek and Turner (2004)...	83
Figure 2.27: CTOSE Phases of Response, from: Broucek and Turner (2004) (Contents SIC) .....	83
Figure 2.28: IT Security Fundamentals, from: Jeong (2006) .....	85
Figure 2.29: Digital Forensics Investigation Fundamentals, from: Jeong (2006) .....	85
Figure 2.30 Process flow between the roles in digital forensics investigation, from: Jeong (2006) .....	87
Table 2.4: The six questions applied to all participants in FORZA, from: Jeong (2006) .....	88
Table 3.1: “Basic phrase structures of English”, from: McMenamin (2002) .....	100
Table 3.2: “Commonly used transformations of English”, from: McMenamin (2002) .....	100
Figure 3.1: The negative transformation in the word “Unlikely” .....	101
Table 3.3” “Linguistic levels in spoken and written language”, from: McMenamin (2002, Section 2.1.2) .....	102
Figure 3.2: “Waveform, intensity contour, and fundamental for English vowels”, from: McMenamin (2002, Section 4.2.2) .....	103
Figure 3.3: “Spectrogram for English vowels”, from: McMenamin (2002, Section 4.2.2) .....	104
Table 3.4: “Spelling of “Maryanne” in QUESTIONED and KNOWN writings”, from: McMenamin (2002, Section 4.2.2) .....	105
Table 3.5: “The science of linguistics”, from: McMenamin (2002, Section 2.2.1) (Contents SIC) .....	108
Table 3.6: “An example of a possibly unique writing style”, from: McMenamin (2002, Section 3.4.1) .....	110
Table 3.7: “Relative proportions of variants for each of the five variables”, from: McMenamin (2002, Section 3.5.2) .....	112
Table 3.8: “Sentence string with seven opportunities for word choice”, from: McMenamin (2002, Section 3.6) .....	112
Table 3.9: “Two of the 36,864 possible choices”, from: McMenamin (2002, Section 3.6) .....	112
Table 3.10: “Examples of Linguistic Norms”, from: McMenamin (2002, Section 6.2) .....	116
Figure 3.4: Sample Image of the Phrase Context Program GUI, from: Mortensen (No Date) .....	117
Figure 3.5: Sample Image of the Concordance Program GUI, from: Concordance (2009) .....	119
Figure 3.6: Sample screens of the T-LAB program GUI, from: TLAB (2009) .....	121
Figure 3.7: Pre-process Panel, from: WEKA (2009) .....	122
Figure 3.8: Classifier Panel, from: WEKA (2009) .....	122
Figure 3.9: Pre-process Panel, from: WEKA (2009) .....	123
Figure 3.10: Associate Panel, from: WEKA (2009) .....	124
Figure 3.11: Select Attributes Panel, from: WEKA (2009) .....	125
Figure 3.12: Visualize Panel, from: WEKA (2009) .....	126
Figure 3.13: Interactive decision tree construction Panel, from: WEKA (2009) .....	127
Figure 3.14: Neural Network GUI, from: WEKA (2009) .....	128
Figure 3.15: Wmatrix GUI, from: Rayson (2008) .....	129
Figure 3.16: Wmatrix Tag Wizard, from: Rayson (2008) .....	130

Figure 3.17: Wmatrix View Folder Function, from: Rayson (2008).....	131
Figure 3.18: Wmatrix Concordance Function, from: Rayson (2008).....	132
Figure 3.19: Wmatrix Compare Frequency Lists Function, from: Rayson (2008)....	133
Table 3.11: “Definitions related to variation, individualization, and writings”, from: McMenamin (2002, Section 6.4) (Contents SIC) .....	134
Table 3.12: “Conclusions on resemblance between questioned and known writings”, from: McMenamin (2002, Section 6.4.5) .....	137
Table 3.13: “Conclusions on consistency within known and questioned writings”, from: McMenamin (2002, Section 6.4.5) .....	137
Figure 4.1: Cyber Forensic Linguistics as a Component in the Greater Realm of Forensics .....	141
Figure 4.2: General Conceptual Framework for Forensic Linguistic Analysis .....	144
Figure 4.3: Similarities between Klopper’s framework for problem-solution oriented research and the general conceptual framework for forensic linguistic analysis.....	147

## *Chapter One*

### PROBLEM STATEMENT AND RESEARCH DESIGN

*A general principle underlying ethical writing is the notion that the written work of an author, be it a manuscript for a magazine or scientific journal, a research paper submitted for a course, or a grant proposal submitted to a funding agency, represents an implicit contract between the author of that work and its readers.*

Roig (2006: 2)

#### **1.1 Introduction**

This study presents a conceptual framework of cyber forensic methods to help forensic auditors to differentiate between ethical academic writing and its negative counterparts, plagiarism and other forms of idea theft. This will incorporate forensic linguistics into cyber forensics in order to extend cyber forensics beyond computer hardware and software. The proposed new method makes available a set of grammatical and semantic criteria to help determine contraventions of ethical writing. A combination of forensic auditing systems may be necessary in instances where no single technique on its own would provide conclusive evidence of contravention of the principles of ethical writing.

In the following three sections, the central concepts of this study are briefly outlined. Thereafter the need to conduct a study of this nature is motivated, and the problems under investigation, the aim and objectives of the study, and the research questions to be answered at the end of the study are presented.

#### **1.2 The Term “Conceptual Framework”**

The term “conceptual framework” is not well understood. According to the organisation Mujer Sana (2003), a framework enables the user to better explain why something is being done in a particular way. The author/s go on to state that a framework caters for the use of ideas and systems developed by other people, as well

as enables the user to explain why certain steps were taken and what outcomes the user would like to achieve. Conceptual frameworks are used in diverse fields. “Conceptual framework” is a fairly new concept within the discipline of ethical writing. One can therefore postulate that a conceptual framework is a pre-empirical framework in contrast with a model which takes as its input empirical data to determine the predictive power of the model.

This study sets out to construct a conceptual framework that locates ethical academic writing and its counterpart, unethical academic writing, within the greater domain of forensic auditing. This conceptual framework, once constructed, will lend itself to further use within the field of corpus stylistics.

### **1.3 The Term “Ethical Writing”**

According to Roig (2006), the concept of “ethical writing” plays an important role in the creation of academic and other professional documents. He further explains that ethical writing constitutes an implicit contract between authors and readers. A breach of ethical writing can generally be attributed to commitment of plagiarism. He considers plagiarism to be one of the most serious violations of the above-mentioned contract. Plagiarism is one of the main forms of unethical writing. Ethical writing is a fundamental requirement for maintaining good quality in academic writing. The term ethical writing is dealt with in greater detail in Chapter 2.

### **1.4 The Relationship between Ethical Writing and Cyber Forensics**

The terms “cyber forensics” and “computer forensics”, are close cousins, as can be seen in Zatyko (2007), who states: “Computer forensics, also called cyber forensics, is the application of computer investigation and analysis techniques to gather evidence suitable for presentation in a court of law. The goal of computer forensics is to perform a structured investigation while maintaining a documented chain of evidence to find out exactly what happened on a computer and who was responsible for it.” This definition shows that computer forensics and cyber forensics are simply different terms describing the same thing. Yet another similar term to take into account is that of “digital forensics”. This is generally understood to mean the same thing as cyber and computer forensics, even though a written definition may be very difficult to come by. Jeong (2006) *cites* Pollitt (2004), who states that there is no single definition for the term “digital forensics.”



## 1.5 Motivations for the Study

Conversations with fellow students and with academics reveal that the relationship between the concepts “ethical writing, unethical writing, plagiarism” and “fraud” seems to be not well understood, which would make the problem under investigation an epistemic one.

A *Google* search, using the term “ethical writing” on 19/11/2008, returned 8,200 positive results constituting online tutorials or downloadable documents about the topic. Of these, a number seem to be useful in spite of the fact that such material has not been tested through the process of peer review.

On the same day a search of the online scholarly database, *Science Direct*, using the same search term, “ethical writing” for articles that have appeared in academic journals, returned zero results. In spite of the manifest ignorance regarding the topic, as evidenced in the absence of peer reviewed literature in scholarly journals, ethical writing is important because it is the positive side of the coin for which forms of unethical writing like plagiarism and fraud constitute the negative side.

The problem of sparse literature regarding ethical writing and its counterpart, unethical writing, along with an apparent poor understanding of the complex relationship between the elements of ethical and unethical academic writing, indicate that the construction of a conceptual framework for the comprehensive study of the problem-set is warranted and that an empirical study is premature.

Many institutions are plagued by the issue of plagiarism, which is an ever growing problem thanks to advances in technology and increased knowledge in the usage of computers. In South Africa, plagiarism is something that academics and institutions do not understand well. There is an imperative need to identify and bring to light the various forms of plagiarism and intellectual property theft. Intellectual property is defined as “intangible property that is the result of creativity, e.g. patents or copyrights” (AskOxford.com, 2008). In order to have an effective anti-plagiarism system in place, good ethical writing practices need to be put into effect and adhered to. Ethical writing is a factor that underpins all written texts, more especially ones that are aimed at publishing. This research aims to explain the concept of ethical writing as well as delve into what constitutes breaches, and various methods to combat unethical writing.

There are many forms of forensics in existence today, for example, forensic ballistics, forensic pathology, forensic psychology and forensic linguistics. One of the newer forms is that of “cyber” forensics: this term has come about thanks to the boom in the computer industry and pertains solely to electronic devices and media. There are other types of forensics that pertain to the same phenomenon. These are known as digital and electronic forensics. However, this link has yet to be formally established, which this research aims to achieve. Forensic stylistics is yet another way in which authorship can be determined and is a factor commonly overlooked when searching for plagiarism. This research thus aims to highlight the possible relationship between stylistics and plagiarism, as well as to bring to light the forms and methods that are prevalent in committing electronic fraud.

### **1.6 Problem Statement**

Having provided a general introduction and motivation for the proposed thesis in the previous sections, this section will present a formal statement of the problem-set that warrants this study. Because ethical writing is a concept that is not fully understood, guidelines need to be developed as to how a writer can adhere to good ethical writing principles, and to help identify breaches in ethical academic writing. Furthermore, plagiarism, one aspect of unethical writing, is not always easy to identify, as shown in Chapter 2. There are many different types of plagiarism, which make it very important for the reader to be knowledgeable about the subject and well-versed on the various forms of idea theft. With the wealth of information available on the Internet, specific tools, e.g. *Turnitin.com* or *My Drop Box*, are commonly used to aid in the detection and prevention of plagiarism. However, the effectiveness of these tools still needs to be demonstrated, as it would be a waste of time and money to use something only to find out later on that it does not live up to the claim of accurately identifying instances of plagiarism. For example, many of these tools have a repository containing written works from hardcopy journals that have been scanned into electronic format. Should material be plagiarised from a source that has yet to be scanned into this database, there is a very strong possibility that the offence would go undetected.

Once tools and techniques to prevent, deter, detect and diagnose plagiarism have been devised, legal factors must then still be taken into account. This is to ensure

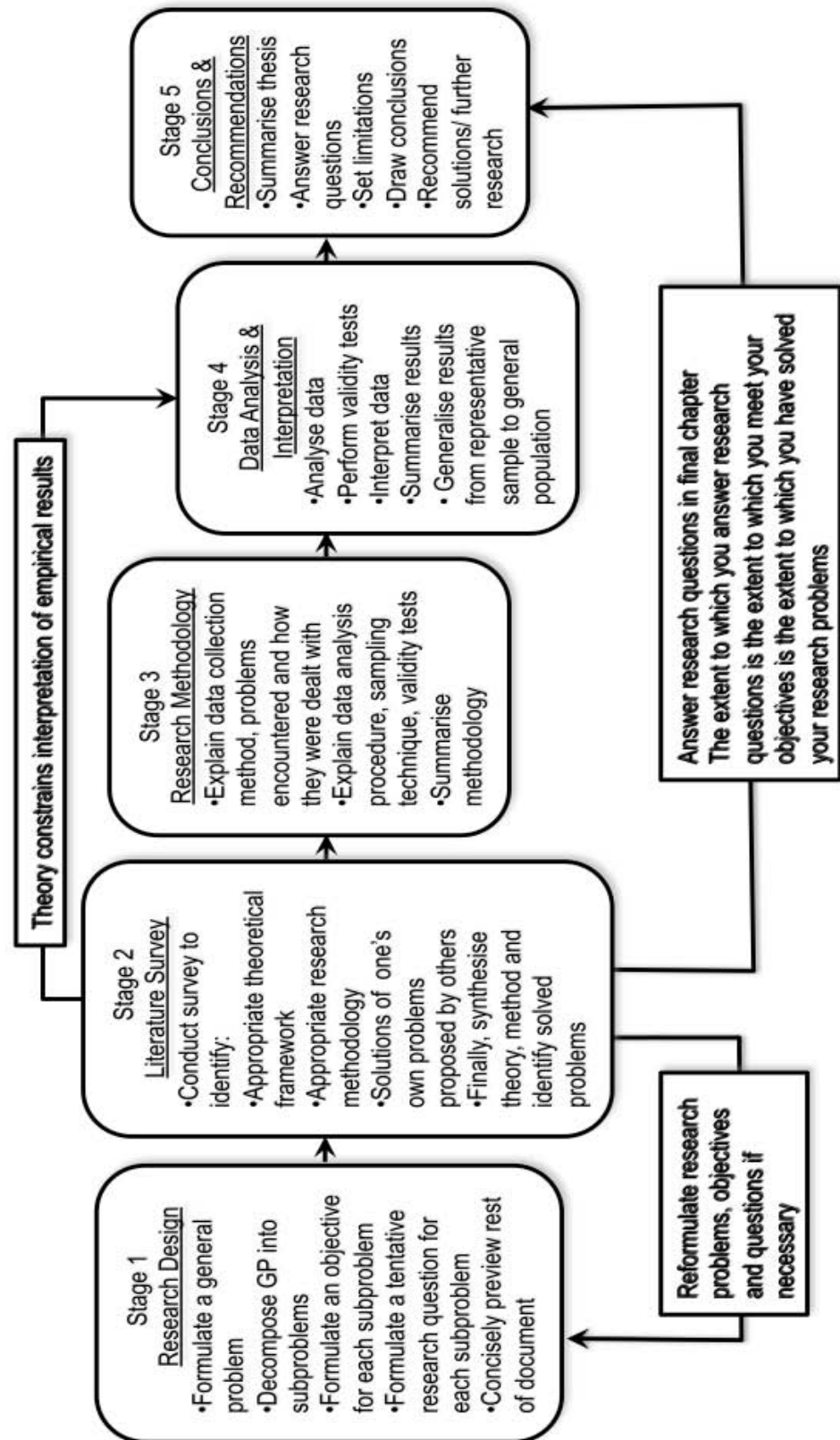
that the process does not violate the suspected plagiariser's individual rights. If an investigation into suspected plagiarism does not comply with the legal requirements, such an investigation will not stand up in a court of law.

Another aspect that has yet to be established is the contribution of forensic linguistics as part of forensic auditing to determine plagiarism. Forensic linguistics involves the scrutiny of auditory and written texts, based on specific linguistic principles (relating to the phonetic, philological, morphological, syntactic and stylistic aspects of language use) to determine patterns or deviations in the norms. This could prove very useful in the fight against plagiarism as it would give users another dimension with which to detect this ever growing problem. Forensic linguistics methods in conjunction with other anti-plagiarism tools, used within a coherent cyber forensic framework, could form part of an effective strategy against plagiarism and intellectual property theft.

The thesis follows Klopper (2008)'s conceptual framework for problem-solution oriented research, which indicates that relevant research begins with the identification of a problem to be solved, and ends with the researcher's self-assessment of the extent to which s/he is able to answer the research questions in order to meet research objectives, thereby solving the problems that motivated the study in the first place:

## Problem-Solution Oriented Research

This schematic outlines the phases that are required in **problem-based research** to provide credible empirical results, and to achieve **coherence, progression and closure**.



Rembrandt Klopper 2008©

Figure 1.1: Framework for problem-solution oriented research, from Klopper (2008) (Contents SIC)

In this section, the general problem under consideration will be formulated and decomposed into subproblems. The general problem can be stated as follows: there does not yet exist a cyber forensic framework that would accommodate the roles that various forms of text analysis could fulfil as part of cyber forensic auditing.

### **1.6.1 Subproblem 1**

It has not yet been determined what roles various forms of text analysis could play as forensic tools in determining the quality of ethical academic writing.

### **1.6.2 Subproblem 2**

There is not yet a conceptual framework for cyber forensic auditing that accommodates the study of ethical writing as part of cyber forensics.

The study will focus on solving these two subproblems. Subsequent research will focus on constructing a framework for ethical academic writing that due to the nature of ethics, falls outside the scope of a masters thesis and because the construction of a framework for ethical academic writing requires that the relationship between forensic linguistics and cyber forensics first be clarified.

## **1.7 Aim and Objectives**

The general aim of this study is to ascertain the requirements needed to ensure that the quality of academic writing remains at an optimal level. The research will focus on aspects of plagiarism, cyber forensic methods, stylistics and other forms of intellectual property theft. This aim can be decomposed into the following specific objectives:

### **1.7.1 Objective 1**

To determine what roles various forms of text analysis could play as forensic tools in determining the quality of ethical academic writing.

There are several tools and techniques available that could play a part in assessing the quality of ethical academic writing. These tools, for example *Turnitin.com*, need to be analysed and benchmarked for the specific purpose of determining the quality of ethical writing. Another factor to take into account is that certain tools could complement each other and open up other avenues of analysis.

### **1.7.2 Objective 2**

To establish a conceptual framework for cyber forensic auditing that accommodates the study of ethical writing as part of cyber forensics.

The system of a conceptual framework will allow the researcher or any other user to quickly identify the various possible routes that exist for a particular problem when going about an ethical writing investigation.

### **1.8 Interim Research Question and Subquestions**

In this section, interim research questions are presented in view of the fact that according to Klopper (2008), the literature review phase of research may reveal that other researchers have solved some of the problems identified by the present researcher, requiring the researcher to revise the initial problem statements, objectives and research questions. This entails that the status of the before-mentioned aspects of research design could be reformulated post the literature review.

#### **1.8.1 General Research Question**

Could intellectual property theft constitute a breach of ethical writing?

#### **1.8.2 Subquestion 1**

What roles could various text analysis techniques play as forensic tools in determining the quality of ethical academic writing?

#### **1.8.3 Subquestion 2**

What would be the elements of a conceptual framework for cyber forensic auditing that accommodates the study of ethical writing as part of cyber forensics?

### **1.9 Envisaged Contribution to the Body of Knowledge**

No official research exists regarding the relationships between stylistics, plagiarism and ethical writing, nor has a relationship between cyber, digital and electronic forensics been established, yet a relationship does exist, which is what this study aims to demonstrate. This research seeks to uncover the links between all the above mentioned factors, thereby adding to the body of knowledge. A by-product of

the research will show the effectiveness of certain online tools in detecting plagiarism. The research will also give the reader a tested and accurate system to detect, diagnose, prevent and prosecute plagiarism.

### **1.10 Research Design**

The research will be conducted in a pre-empirical qualitative framework analysing data already in the open domain, and will therefore not contain an empirical phase during which primary data is collected and analysed. It will compare various approaches to ethical writing in order to construct a comprehensive and coherent framework for detecting breaches of the principles of ethical writing. The project will identify modes of linguistic analysis such as grammar, syntax, word frequency analysis, and stylistics to help establish authorship of documents, and will also identify online tools and downloadable software programs to help determine the extent of plagiarism within academic assignments. Methods will be cross-checked and contrasted against one another to clearly illustrate the strengths and weaknesses of the various approaches. The intention of the study is to identify dependable means by which unethical academic writing such as plagiarism, either intentional or unintentional, can be established.

### **1.11 Layout of the Study**

Chapter 1 – Problem Statement and Research Design

Chapter 2 – Literature Review

Chapter 3 – Forensic Linguistics

Chapter 4 – The Research Questions Revisited

Chapter 5 - Conclusions

### **1.12 Conclusion**

This chapter provided a brief outline of the study to be undertaken and offers an understanding of the purpose, intended objectives and research design of the study. It also gave an understanding of the pertinent concepts of ethical writing and forensics. The research design and reason for the study were also briefly outlined. The following

chapter (Chapter 2) will discuss the concepts of e-forensics, stylistics, linguistics and frameworks as well as explain any models that may apply to the study.



## *Chapter Two*

### LITERATURE REVIEW

#### **2.1 Introduction**

According to Broucek and Turner (2004), in the case of computer misuse and e-crime, it is the evidence acquisition step that presents the most difficult technical and legal challenge in terms of e-Forensics. The study aims to bring together the legal aspects and tie them in to the process through which the investigation should be done. e-Forensics is a specialised field, as the crimes committed on computers are usually more difficult to notice due to the fact that they can be concealed with far more ease. This being said the type of people most suited to this line of work are those with experience in computer security, programmers and those with hacker-like abilities as they are best able to identify and track down computer misuse. However, nowadays with the rapid expansion of the Internet, as well as information being made more readily available, it is often the case that students are the ones guilty of intellectual property theft, i.e. plagiarism.

As indicated in the previous chapter (Chapter 1), the research aims to identify the various types of plagiarism in use, how to detect them and explain the usage and benefit of electronic anti-plagiarism tools as well as provide an understanding of stylistic analysis. Once these factors are ascertained, a close study of the nature of electronic assignments will be done. The aim of this will be to expose weaknesses and suggest solutions. e-Forensics in academic assignments is a very new field, as there are not many techniques or tools in existence to combat misuse of academic assignments. However, one must question how many cases slip past before a person does get caught; no doubt a system needs to be in place to detect and expose this kind of offence.

#### **2.2 Cyber Forensics and the Nature of Electronic Material**

Electronic material is in almost all cases capable of being copied. Since this study deals mainly with academic writing, the focus will be mostly on the copying of text specifically. Adobe ([www.Adobe.com](http://www.Adobe.com)) allows documents to be digitally signed and various aspects protected, making it more difficult to simply copy and paste from

such documents. A version of Adobe known as Adobe Professional contains a variety of features such as Scanning, OCR and text editing among others. If an Adobe document is not password-protected properly, a user may manipulate the document in any number of ways. It is, in general, a very simple process to copy protected text from any electronic document. One of the most extreme examples would be a simple matter of printing out the protected material and then scanning it back into the computer using optical character recognition software (OCR software), thereby rendering the document editable. This is not a very difficult task and this being said, it is clear that academics and institutions should be aware of the ways in which plagiarism can be committed in order to better limit and eventually eliminate the problem.

Marcella and Robert (2006) define cyber forensics as the process of extracting data from computer storage media and ensuring its validity and reliability is not compromised. The importance of the accuracy of the extracted data plays a pivotal role in the admissibility of evidence in a court of law. According to the Information Systems Audit and Control Association, hereon referred to as ISACA (2004), computer forensics can be defined as a process of extracting data and information from electronic storage devices using court validated tools and processes in order to maintain accuracy and reliability of the evidence. It is clear from these definitions that one of the primary concerns of computer forensics is to ensure the evidence is admissible in a court of law and does not get compromised in any way during the course of the investigation. ISACA (2004) adds that the main challenge facing the investigator is finding, collecting, preserving and presenting the evidence and data in a court of law. According to ISACA (2004), computer forensics is a science and an art, in the sense that data have to be acquired from a source in order to determine if a crime has occurred. Organisations that have good security systems and keep detailed logs do not face much difficulty when a computer crime occurs. The terms digital forensics and electronic forensics also amount to the same thing. They all deal with the acquisition of electronic data from computer storage. In a similar light, Carrier and Spafford (2004) explain the forensic investigation system to be a combination of science and technology that is used to develop theories regarding possibilities of certain outcomes that can be used in a court of law. They further state that a digital forensic investigation is a similar process, except that in this instance, science and

technology is used to examine digital objects. There are factors which could trigger a computer forensic investigation and it is up to the IS auditor to brief the organisations on the signs and procedures (ISACA, 2004). According to ISACA (2004), computer forensics can be used in the following situations: whistle blowers, HR investigations, fraud investigations and compliance investigations.

There are also several different types of analysis depending on the type of digital data in question. Carrier and Spafford (2004) list a few examples of different digital analysis types. Media Analysis is a form of analysis which only looks at data stored on a storage media type, for example external hard drives. Media management analysis is a form which looks at systems that are used to organise media; an example here would be RAID systems. File system analysis pertains to the analysis of a file system from inside a partition. Application analysis deals with the analysis of data within a file, often application specific. In network analysis, data on the network can be analysed, such as packets. Operating System analysis, output and configuration files can be looked at. Executable analysis is a form of analysis of objects that cause actions to occur. Image analysis pertains to the analysis of digital images as, for example, a person could be added or removed from a picture thereby altering courses of action. Video analysis deals primarily with the analysis of video footage from security cameras. All these types of digital analysis deal with different types of electronic material; it is therefore essential to have a good understanding of the various forms in order to facilitate an effective investigation process.

### **2.3 The Role of Language Analysis in Forensic Auditing**

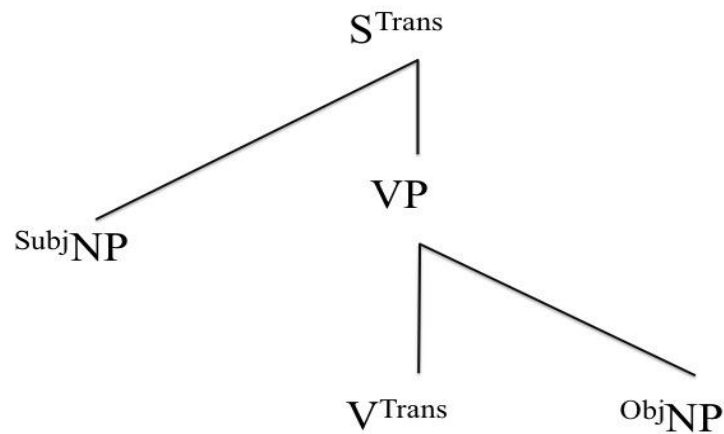
In view of the fact that the primary resource of plagiarism is language in the form of texts, built up from words that are combined according to language rules into sentences and paragraphs, an investigation of the forensic aspect of plagiarism needs to determine the role that particular aspects of linguistics could play in forensic analysis, for instance, the use of sentence patterns (syntax) and lexical patterns (lexicography and semantics).

One of the problems around involving linguistics in forensic analysis is that this discipline has been extremely productive over the past century, which has led to it developing a variety of schools of linguistics, such as Descriptive Linguistics, Functional Linguistics, Relational Grammar, Generative Grammar, Sociolinguistics,

Psycholinguistics, Computational Linguistics, and Cognitive Linguistics, to name but a few (Holcombe, 2007). The following section uses Quirk et al., (1974) as point of departure.

### 2.3.1 Using Syntactic Patterns in Forensic Analysis

Following Quirk and his fellow researchers, sentences can be characterised as generic lexical categories that are hierarchically linked in particular sequences that preserve the correct order of particular words in sentences. A graphical representation of such sentence patterns is known as a tree diagram. Tree diagrams reveal that a sentence is built up of immediate constituent phrases like a noun phrase (NP), a verb phrase (VP), and a propositional phrase (PP):

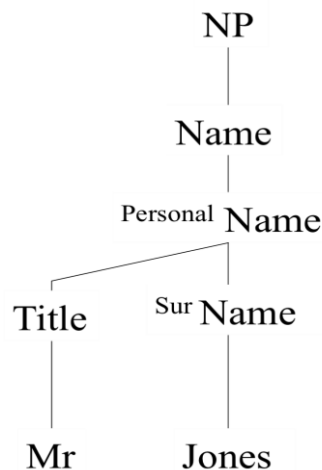


**Figure 2.1: The general pattern of English transitive sentences**

The above diagram (Figure 2.1) depicts a typical transitive sentence that could be broken down into a noun phrase (NP) and a verb phrase (VP). The verb phrase is then further parsed (broken down) into its immediate constituents, the transitive verb and its object. The tree diagrams that follow represent the different types of NPs from among a greater variety of English NPs, namely a NP of which a name is the nucleus, also known as the core constituent. The <sup>Subj</sup>NP is the peripheral or satellite constituent. It can therefore be said that phrases have a nucleus-satellite (nuk-sat) setup. Furthermore, it can be said that a NP is a phrase that has some type of a noun, a name, a pronoun or a proper noun as nucleus (core constituent), which can be complemented

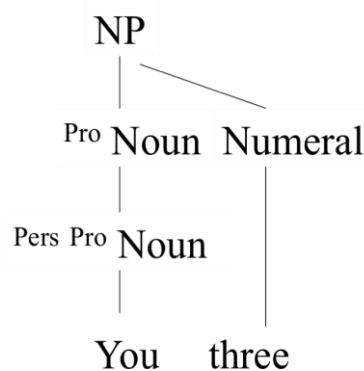
by a range of satellite (peripheral) constituents such as titles, numerals, adjectives and articles.

The three tree diagrams that follow represent three different types of NP from among a greater variety of English NPs, namely a NP of which a surname is the nucleus (core constituent), followed by a second NP diagram in which a personal pronoun is the nucleus (core constituent), and finally, the diagram of a complex NP with a range of pre-determiners, with a proper noun as its nucleus (core constituent):



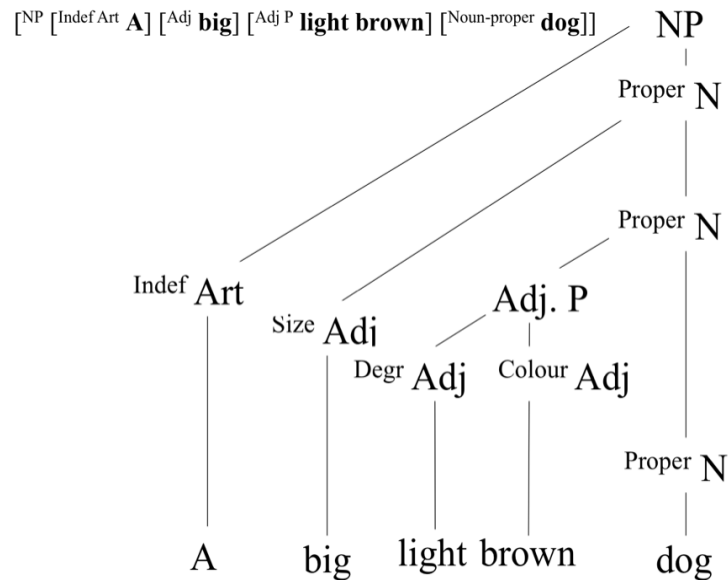
**Figure 2.2: A typical English noun phrase consisting of a predetermining title and a surname as the core constituent of the phrase.**

Here the NP, which is a name, is broken down into a title and sur name. The surname is the core constituent of the NP.



**Figure 2.3: A typical English noun phrase consisting of a personal pronoun as core constituent and a numeral as post-determining constituent.**

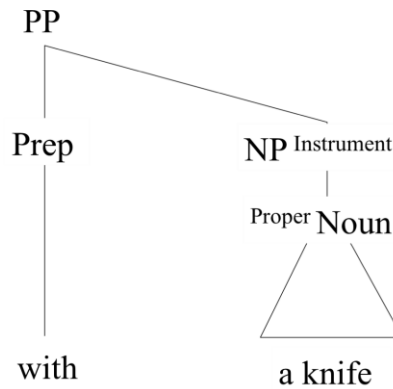
The diagram above depicts the use of a specific personal pronoun and a numeral. Here the personal pronoun “you” and the numeral “three” are combined to form the object of the sentence.



**Figure 2.4: An English noun phrase consisting of a predetermining indefinite article, an adjective indicating size, an adjectival phrase consisting of an adjective indicating degree, and an adjective indicating colour, as predetermining constituents, with a proper name as the core constituent of the phrase.**

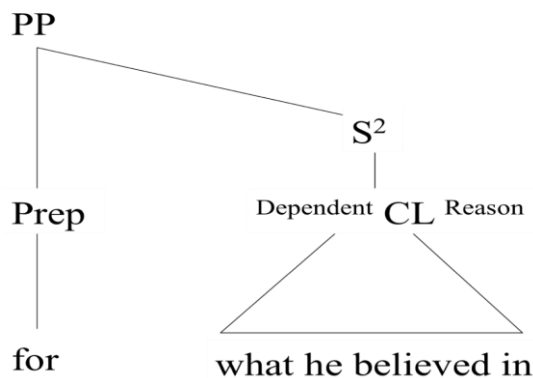
Figure 2.4 depicts a NP with a complex phrase structure. It is important to note that the words within square brackets at the top of the diagram are a form of linguistic annotation that stipulates the combination patterns of lexical categories that make up the hierarchical structure of the NP. This type of annotation is used in many different corpus analysis techniques. The term “corpus analysis” refers to the various forms of text analysis employed to analyse bodies of text and is further discussed in section 2.7.1. The diagram above shows a phrase with an indefinite article “a”, the size adjective “big”, the degree adjective “light”, the colour adjective “brown” and lastly, the proper noun “dog”.

Figure 2.5 below shows a PP (prepositional phrase) with the preposition “with” as the nucleus (core constituent) and the satellite constituent here being the object “knife”.



**Figure 2.5: An English prepositional phrase consisting of a preposition as the core constituent of the phrase, and a noun phrase indicating an instrument as post-determining constituent.**

Figure 2.6 below shows a PP with the preposition “for” as the nucleus (core constituent) with the dependent clause “what he believed in” as the satellite constituent:



**Figure 2.6: An English prepositional phrase consisting of a preposition as the core constituent of the phrase, and a dependent clause as post-determining constituent.**

From the above tree diagrams it should be clear that sentence patterns serve as generic algorithms that allow people to populate them with specific lexical items to form phrases. Due to their generic nature, sentence patterns in themselves have little value in forensic auditing to help establish particular authorship, while the specific lexemes (words) that make up the lexical items can be used to establish a writing profile for a certain individual, as these words are dependent on a variety of factors specific to each individual writer (as will be demonstrated in the sections that follow).

### 2.3.2 Using Word Patterns in Forensic Analysis

In this section, the focus shifts from form to meaning as a possible tool in forensic auditing. The diagrams in this section depict the implied relationships

between words, based on shared attributes of meaning that bind them together as related words. This section serves two functions, namely, to introduce a number of the key concepts of this study, and to demonstrate how humans assign meaning to words through the use of other words that are in turn defined by means of other words. The vast network of words associated with one another in this way combine to give meaning to human language. This process is known as semiosis and it is important to note that semiosis brings an element of indeterminism to how humans understand one another, often resulting in misunderstandings and alternative interpretations of intent. According to Eco (1980: 1), signs and occurrences can only give specific implied meaning based on the context in which they are used. Signs, in this context, are taken to mean the way in which the reader gives meaning to the text; one cannot read text and not derive meaning from it. Eco presents a concept known as the “hermeneutic circle” describing the sign interpretation as type of “bet” in relation to how the reader will interpret the sign, a system whereby the odds of this “bet” can be swayed by the context in which the sign is used. This allows the writer to coax the reader into drawing up meaning based upon the context in which the words are used (Dutton, 1992). Eco (1994: 2) goes on to discuss the “semiotic web”, a concept that includes both the study of signification systems (language usage) and the processes by which users apply these systems for purposes of communication (Eco, 1994: 2). These concepts play a pivotal role in this study as they show that words and meanings are derived from other words.

It is this relationship between words that are of importance to the forensic linguistic investigator, as each individual has a unique word usage and sign interpretation. The corpus analysis process encompasses these factors when performing word frequency analysis as is evident in both function word and concordance analysis. Eco’s views in the previous paragraph are the underlying systems behind the modern forms of word frequency analysis.

Word usage is a key concept in cyber forensics when determining authorship. The following diagrams further the explanation of word relationships. The coloured dots in the “spider diagrams” below, formally known as “nodes”, indicate related words and are coloured according to the lexical categories that each word belongs to. The red nodes indicate nouns. A green node indicates a verb. A yellow node indicates



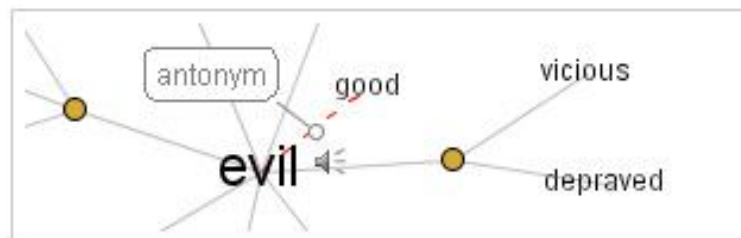
an adjective. A red dotted line indicates words with opposite meanings. Dotted lines show that those words are a form of the parent word.

The spider-grams below follow a standard set of meanings and the key below can be used to understand the various coloured dots and lines. The following diagrams are adopted from [visualthesaurus.com](http://visualthesaurus.com) (2009).



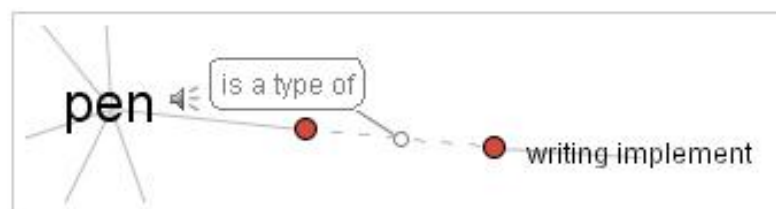
**Figure 2.7: Colour coding for main lexical categories, adopted from: [visualthesaurus.com](http://visualthesaurus.com) (2009)**

A red dotted line indicates an inferred antonymic relationship between lexemes. A loudspeaker image provides a link to the pronunciation of the word.



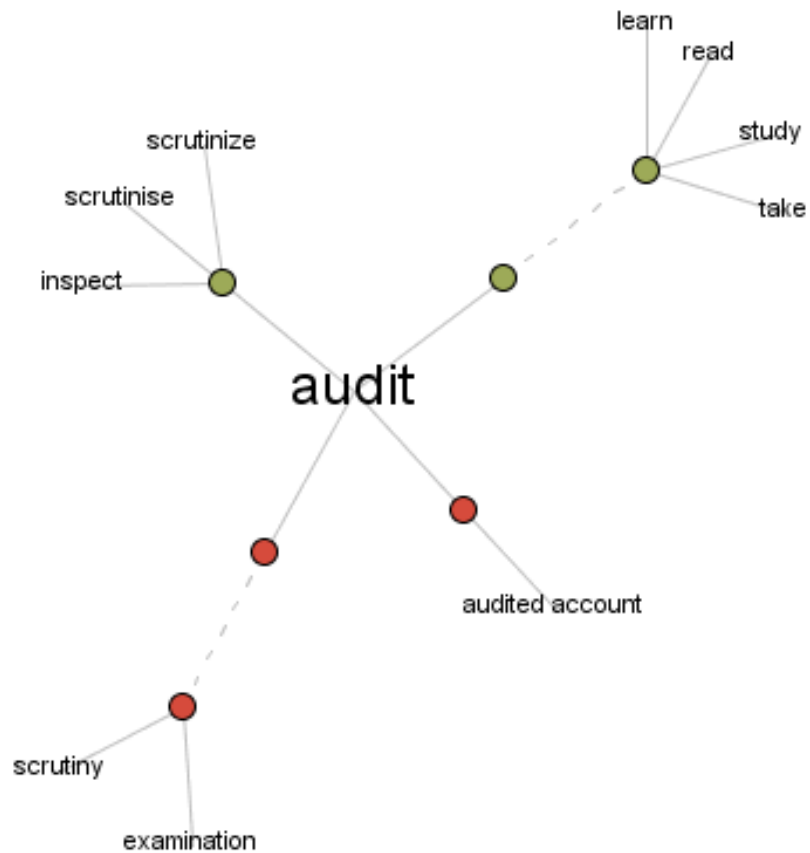
**Figure 2.8: The adjectival antonyms “evil” and “good”, adopted from: [visualthesaurus.com](http://visualthesaurus.com) (2009)**

The diagram above depicts an antonymic as well as two synonymic relationships. It is clear that “good” and “evil” are opposites and reflect opposite meanings.



**Figure 2.9: “Pen”, adopted from: [visualthesaurus.com](http://visualthesaurus.com) (2009)**

This diagram depicts a situation where two nouns are connected by a broken line. Here the grey broken line indicates the lexeme “pen” is one type (subcategory) of writing implement (along with other unnamed writing implements like pencils, paint brushes and stylises).



**Figure 2.10: “Audit”, adopted from: visualthesaurus.com (2009)**

The term “audit” plays a vital role in investigations. As shown in diagram 2.10 above, audit has many verb forms. The words “study”, “read”, “scrutinise” and “inspect” are the functions that encompass auditing. It is clear that this term “audit” and that of “analysis” below, are very closely related, as analysis delves into the investigating and analytic aspects.

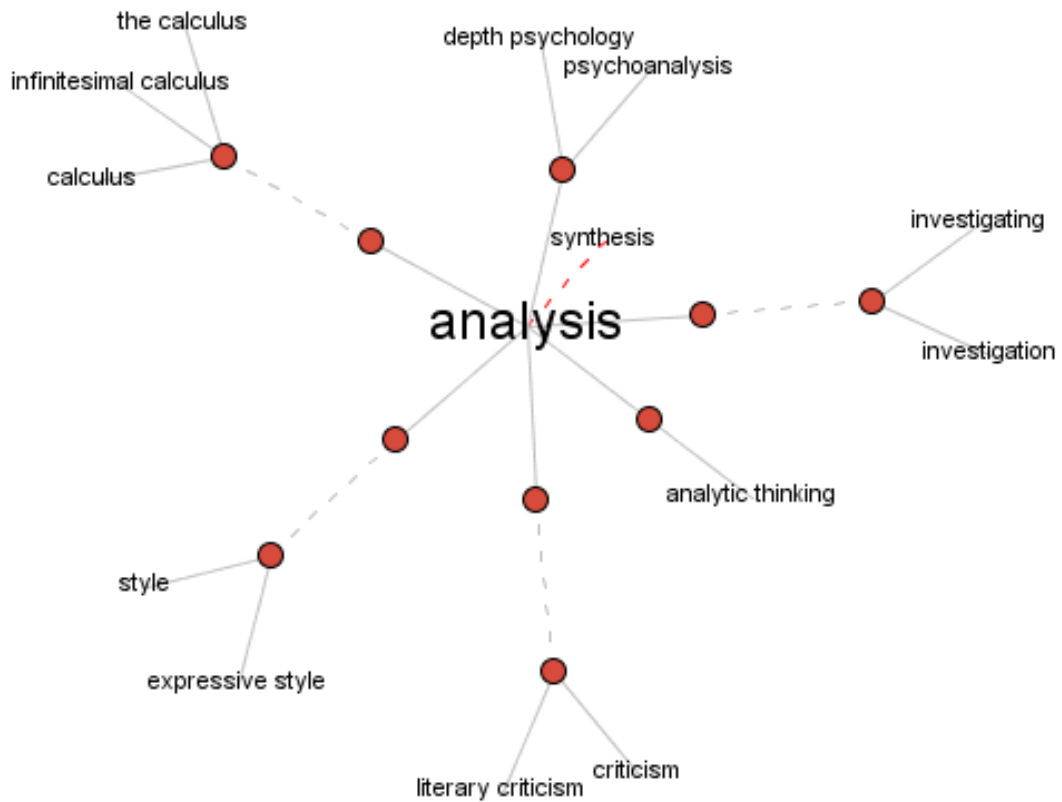
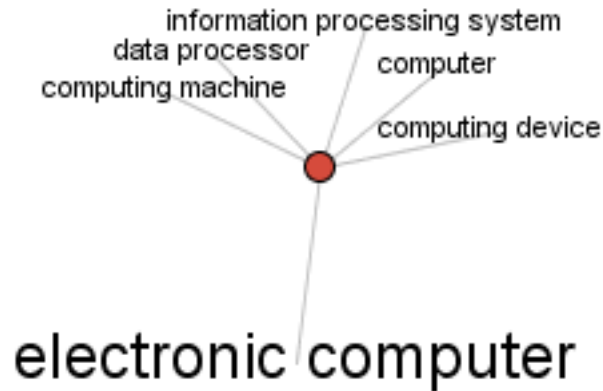


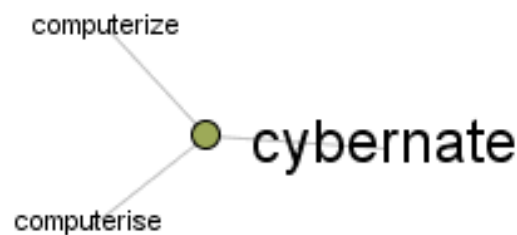
Figure 2.11: “Analysis”, adopted from: [visualthesaurus.com](http://visualthesaurus.com) (2009)

The diagram above depicts many uses for the word “analysis”; for relevance to this study the investigation, analytics and style all play an important part in the cyber forensic investigation. It is also interesting to note that calculus is a function of analysis, giving the notion that analysis involves mathematical aspects.



**Figure 2.12: “Electronic Computer”, adopted from: visualthesaurus.com (2008)**

Figure 2.12 illustrates that there are many concepts that fall into the category of “electronic computer”. Of special interest is that of the words “computer” and “computing device”, as the diagram below shows: in order to “cyberenate” something, one must essentially computerise it. It should also go without saying that “cyberenate” and “cybercrime” have the common element of “computer” in them, except that cybercrime has the crime factor in the mix.



**Figure 2.13: “Cyberenate”, adopted from: visualthesaurus.com (2008)**

Figure 2.13 illustrates that the concepts of “computerise” and “cyberenate” are in fact linked. As mentioned, in order to “cyberenate” something one must “computerise” it. The green dot linking the aspects together represents a verb. It is clear from this diagram that the concepts of “electronic computer” and “cyberenate” are indirectly interrelated via concepts that have the same or similar meaning. Therefore, one can assume that these two concepts are also related to cybercrime.

When one discusses the issues of ethical writing, the above factors as well as those of ethics and writing need to be taken into account. It is clear from the above two diagrams that cyber crime is a form of computer/electronic offence.

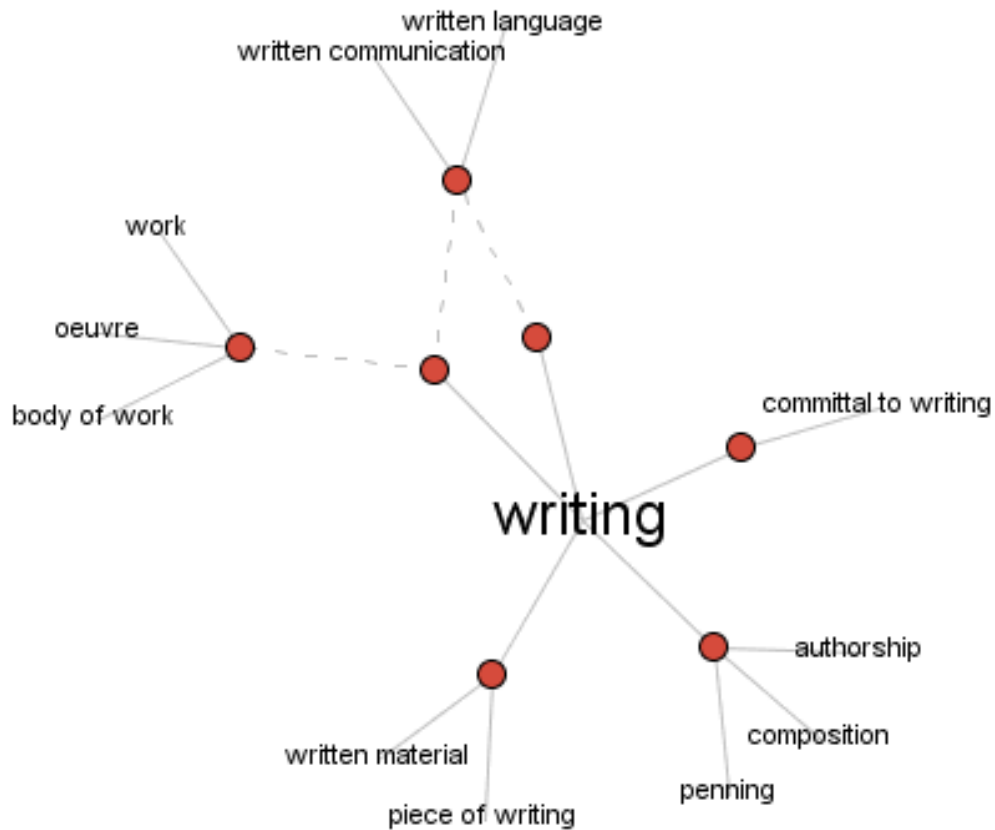


Figure 2.14: “Writing”, adopted from: visualthesaurus.com (2008)

The diagram above (Figure 2.14) depicts the concept “writing” and the subcomponents that play a role in how people understand it. The words on the dotted line section are a direct part of the concept “writing”. Of special importance is the concept of authorship, and the link between this concept and that of writing is clear from this diagram. It also shows that ownership of work plays an important part in the concept of writing.

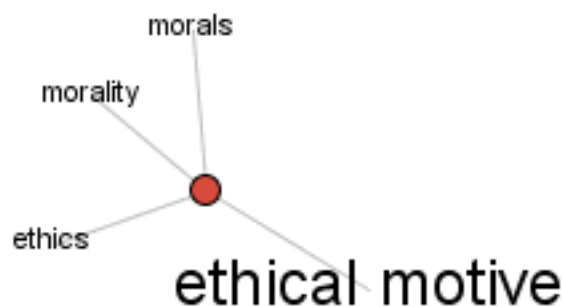


Figure 2.15: “Ethical Motive”, adopted from: visualthesaurus.com (2008)

When attempting to understand the discipline of ethical writing one needs to understand what the concept “ethics” means. According to the diagram above (figure 2.15), ethical motive contains aspects of morality and ethics. This plays an important role in ethical writing as morality affects a person’s decision to knowingly plagiarise another work. The diagram below represents a more clear view of what ethics comprises.

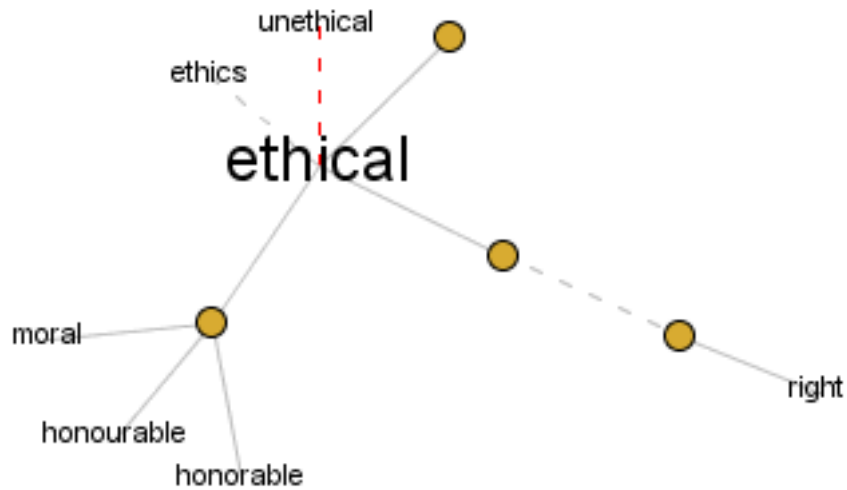


Figure 2.16: “Ethical”, adopted from: [visualthesaurus.com](http://visualthesaurus.com) (2008)

The red dotted line in the diagram above (figure 2.16) depicts the opposite of the word in question, in this case, the concept “unethical”. The word “right” is used to symbolise the outcome of an ethical action in this respect. It also depicts the words “honourable” and “moral” as having a part in ethical behaviours.

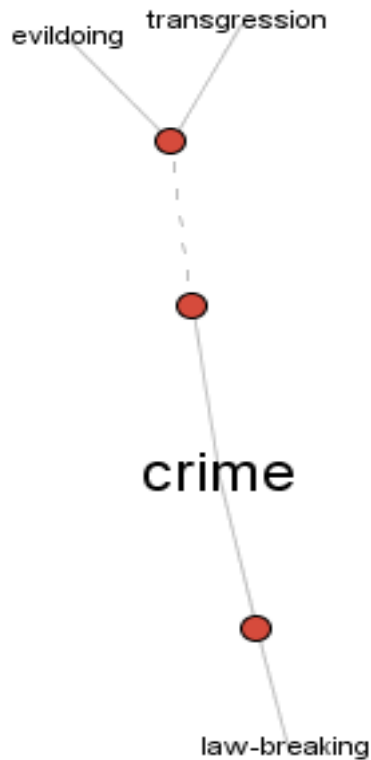


Figure 2.17: “Crime”, adopted from: [visualthesaurus.com](http://visualthesaurus.com) (2008)

The diagram above (figure 2.17) depicts the link between wrongdoing, the law and the concept of crime. It is clear that evildoing and transgression are aspects of the larger concept “crime”. Law-breaking is one of the consequences of crime.

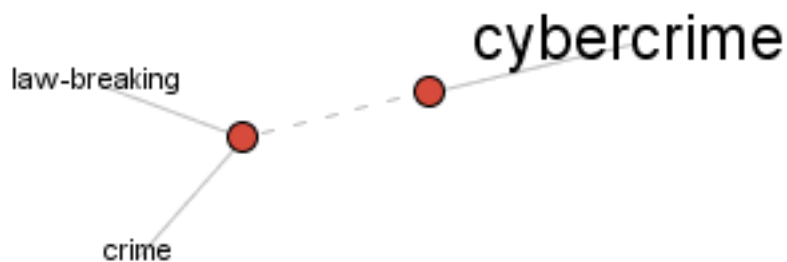


Figure 2.18: “Cybercrime”, adopted from: [visualthesaurus.com](http://visualthesaurus.com) (2008)

Figure 2.18 above depicts the concepts that equate to cybercrime in the electronic realm. Law-breaking, along with crime, which itself embodies any type of illegal activity, is an offence. “Cybercrime” is simply the term used to describe these

two aspects combined. When one looks at the above two diagrams, it is now clear that the concepts of crime and cybercrime are closely interrelated.

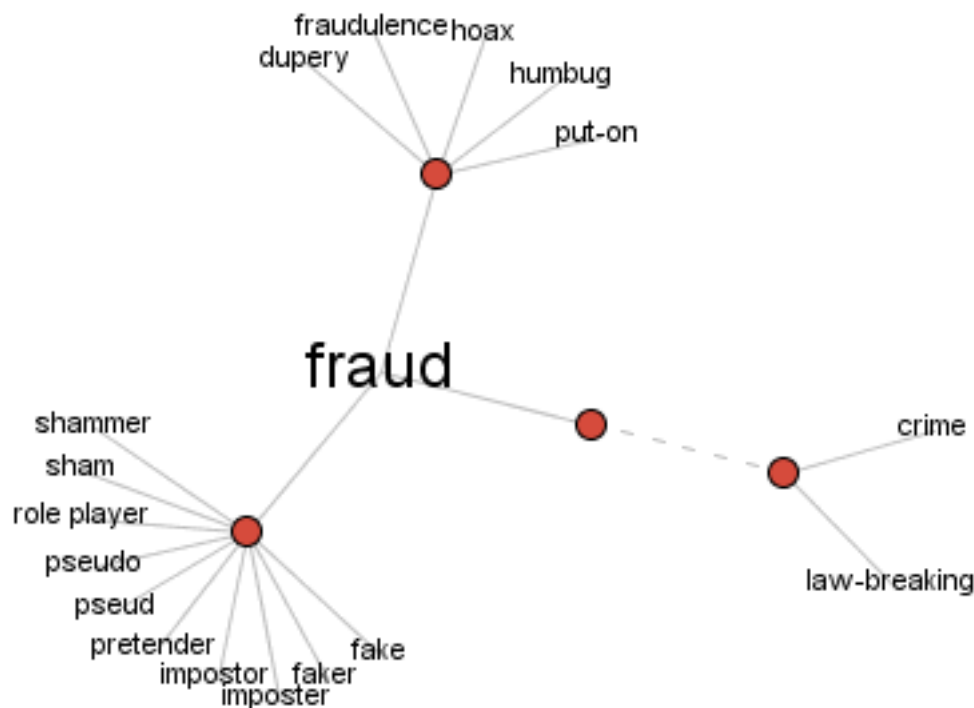


Figure 2.19: “Fraud”, adopted from: [visualthesaurus.com](http://visualthesaurus.com) (2008)

It is important to note in this diagram (figure 2.19), that fraud is associated with crime and law breaking. If we look back at figure 2.18, it becomes clear that cybercrime and fraud are very closely linked as the aspects of “crime” and “law-breaking” are evident in both fraud and cybercrime. From these spider diagrams several important aspects need to be pointed out:

1. We associate words with one another, based on similarities of meaning
2. The meanings of words are defined by means of other words
3. Word association is based on inferences made by language users
4. Some inferences are more tenuous than others, as indicated by the dotted lines in the spider diagrams
5. Part of the association that one has of words, are words that have an opposite meaning as indicated by red dotted lines

From the above five points, one can conclude that forensic researchers cannot depend on the analysis of the meaning of individual words alone to attribute



authorship, due to the inferential nature of language interpretation. Lexical analysis should therefore only form part of a cluster of linguistic tools for attributing authorship.

### **2.3.3 Word Frequency Analysis**

According to Samoilovich (2009), word frequency analysis has many uses. For the purposes of this study, its usage in determining document authorship is of greatest importance. To determine authorship a word frequency count has to be performed on a set of low frequency words (words that are not often used) on a piece of text that is known to be authentic. In simple terms, a word frequency count is the process by which one determines how often a specific word or set of words appear in a specified body of text (corpus). This process is discussed in much greater detail in section 2.7.3 (Function Word Analysis). Once the word frequencies are extrapolated, the same process can be performed on the writing in question and then the results compared. Should the frequency of the writing in question differ vastly from the frequency of that of the writing that is known to be authentic, one can then conclude that the portion of text in question is not original. Samoilovich adds that plagiarism detection is yet another pertinent usage of word frequency analysis. It can allow for the determination of a particular writing style and the determination of authorship without having to go through volumes of data and text, e.g. the Internet.

## **2.4 Ethical Writing**

*“A responsible writer has an ethical responsibility to readers, and to the author/s from whom s/he is borrowing, to respect others’ ideas and words, to credit those from whom we borrow, and whenever possible, to use one’s own words when paraphrasing”*

*Roig (2006: 14)*

Ethical writing is a concept that ties in very closely with the problem of plagiarism. Roig (2006) notes that writing should be characterised by clarity of expression, conciseness, accuracy and most importantly, honesty. Hexham (1999) states that if only 5% of Canadian academics were to plagiarise, it would cost the taxpayer approximately \$200,000,000 a year. It is therefore very clear just how severe

the problem of plagiarism is and how important it is to write ethically. It is also noted that the features of clarity and conciseness can often be transgressed accidentally; this could be due to time pressures or other mitigating factors when completing the written work. However, even an accidental mistake could potentially have serious ramifications. Deliberate misconduct is a very different, as well as a more serious problem altogether. Ethical writing strives to embody the factors that ensure good writing, free of all forms of plagiarism. The factors that embody ethical writing are that it should be clear, accurate, fair and honest (Roig *cites* Kolin, 2002). This statement clearly indicates that in order to adhere to these factors, one needs strong moral grounding. Roig (2006) states that ethical writing is the implied contract between the author of a written work and its readers. He (2006) also notes that there are both intentional and unintentional errors that occur when writing. The unintentional errors come in a few forms, such as unintentional bias. This occurs when the writer, while reviewing the existing data, unintentionally dismisses a particular line of evidence, thereby giving a biased view of the topic. Another form is a factor known as “Cryptomnesia” or unconscious plagiarism; this phenomenon occurs when the writer comes across a concept created by another person and after some time forgets who originally came up with the idea and then believes it to be his or her own. One other common unintentional error occurs when a writer uses much information from another’s work and does not fully credit/reference the source. According to Roig (2006), plagiarism is the most widely recognised and serious violation of trust between the author and its readers. In fact, it is such a serious problem that in some cases offenders have been excluded and dismissed from academic institutions resulting in them losing their jobs. In some instances, academic degrees and qualifications have been rescinded by the universities (Roig *cites* Standler, 2000). The United States Public Health Service has defined plagiarism as one of three major types of scientific misconduct, with falsification and fabrication being the other two forms (Roig *cites* U. S. Public Health Service, 1989).

## **2.5 Referencing**

Proper referencing is one of the most vital aspects in the fight against plagiarism. According to Coxhead (2007), correct referencing will prevent students from being accused of cheating by copying text from other individuals. Coxhead (2007) also adds that referencing is a two step process. The first step is referencing

within the body of a work that the indicated text belongs to another author, and the second step lists the full details of the source further on in the document. There are many different systems of referencing available; however, for the purposes of this study, only the Harvard referencing system will be utilised.

## 2.6 Plagiarism

*“Plagiarism is the deliberate attempt to deceive the reader through the appropriation and representation as one's own the words and work of others. Academic plagiarism occurs when a writer repeatedly uses more than four words from a printed source without the use of quotation marks and a precise reference to the original source in a work presented as the author's own research and scholarship.”*

*Hexham (1999: 2)*

The quotation above is a clear indication of what plagiarism is about. Plagiarism is most likely the most prominent problem associated with written material (O'Connor, 2003). Roig (2006) adds that plagiarism is not restricted to any one discipline. With the advent of more effective word processing software, it has become far easier to “copy and paste” from one source to another with the click of a button. This has made plagiarism a solution for many academic assignments, as noted by students (Barnbaum, 2002). Anti-plagiarism tools do exist, for example, *Turnitin.com*; however, their effectiveness is what needs to be analysed, and plagiarising from the Internet and scanning from books are two different things. This is due to the fact that these anti-plagiarism tools most often search the Internet for similarities in suspected plagiarised work. The situation is very different when it comes to hardcopy books that do not exist in electronic format. If those were being plagiarised then it would be difficult to detect. The repercussions of undetected plagiarism are numerous to say the least. Sometimes plagiarism is straightforward and easy to spot; in other instances, it can be difficult to distinguish as it comes in many forms, not all of which are that apparent.

### **2.6.1 Common Knowledge**

One of the very few instances where referencing and citing sources is not required is a situation where the idea or information is known to be common knowledge. According to Thompson and Olivas (2008), there is much uncertainty even among experts as to what common knowledge is and where its boundaries lie. Two criteria are noted as being most commonly used to determine if something can be regarded as common knowledge: quantity and ubiquity.

- **Quantity**

A fact can be considered to be common knowledge if it can be found in several different sources. For example, Purdue University recommends finding five independent occurrences of the fact in question, before considering it to be part of the common knowledge domain (Thompson and Olivas, 2008 cites Purdue University, 2007).

- **Is the fact Ubiquitous?**

Ubiquitous, with specific regard to this point, means whether the fact is known of everywhere. A fact may be common knowledge in one school, yet it may be unheard of in another. In this scenario, the fact still needs to be properly referenced (Thompson and Olivas, 2008). If the fact is known of everywhere, then it can be assumed that it is common knowledge and therefore will not require referencing.

They additionally note that if one is not sure if something does belong to the body of common knowledge, it is best not to assume that it does. Roig (2006) also reiterates this fact: should a writer not be absolutely certain if the fact in question belongs to the body of common knowledge, then a citation must be provided. In instances such as this, it is imperative that the ideas or works are properly referenced.

### **2.6.2 Types of Plagiarism**

Plagiarism comes in a variety of forms and types, and some forms are very difficult to identify. Being able to detect plagiarism is one of the most important aspects in the fight against this ever-growing problem. This section will establish an understanding of the different forms of plagiarism and provide a means by which each

can be identified. Roig (2006) notes that there are two main forms of plagiarism in existence. These are the plagiarism of ideas and the plagiarism of text.

### **2.6.2.1 Idea Plagiarism**

*“Appropriating an idea (e.g., an explanation, a theory, a conclusion, a hypothesis, a metaphor) in whole or in part, or with superficial modifications without giving credit to its originator...”*

*Roig (2006: 4)*

This statement argues that an idea taken from another individual in any way, needs to be credited accordingly. If not, then it amounts to idea theft. This is a very trivial principle to apply and to avoid, as it entails the proper crediting of any ideas used. Roig (2006) goes on to state that both forms, unconscious as well as deliberate plagiarism, exist in idea plagiarism.

### **2.6.2.2 Text Plagiarism**

*“Copying a portion of text from another source without giving credit to its author and without enclosing the borrowed text in quotation marks...”*

*Roig (2006: 6)*

The statement above is straightforward, as any copying of verbatim (word for word) text is to be enclosed in quotation marks and properly referenced. According to Roig (2006), plagiarism of text is probably the most common form of plagiarism in existence. There are many types of text plagiarism in existence, and the following types are identified by Hexham (1999):

#### **i. Straight plagiarism**

This is best described as the simple removal or addition of words from a piece of text as well as changes in the capitalisation of words. When this process is followed

without referencing, then it is indicative that the offender wished to pass on the superficially changed piece of text as an original work.

The following example (paragraph as well as the reference below it) is taken from Hexham (1999) as examples of plagiarism types:

- **Original Passage**

“But Hertzog recognized the danger and stood up for the rights of the Afrikaner. Only the National Party offered a Christian solution to South Africa's racial problems. The politics of the nationalists, were in the view of *Het Westen*, unquestionably Christian. The Afrikaner People were a Christian people, therefore their politics must of necessity be Christian.”<sup>1</sup>

<sup>1</sup> Irving Hexham, *The Irony of Apartheid* (Lewiston: Edwin Mellen, 1981), p. 185.

- **Plagiarised Passage**

“But General Hertzog recognized the danger and fought for the rights of the Afrikaner. Only the National Party offered a Christian solution to South Africa's racial problems. The politics of the Nationalists, were in the view of the newspaper *Het Westen*, thoroughly Christian. The Afrikaner People were a Christian People, therefore their politics must of necessity be Christian.”

- **Correct Passage**

“Hexham writes “But Hertzog recognized the danger and stood up for the rights of the Afrikaner. Only the National Party offered a Christian solution to South Africa's racial problems. The politics of the nationalists, were in the view of *Het Westen*, unquestionably Christian. The Afrikaner People were a Christian people, therefore their politics must of necessity be Christian.”<sup>1</sup>

<sup>1</sup> Irving Hexham, *The Irony of Apartheid* (Lewiston: Edwin Mellen, 1981), p. 185.

## ii. Plagiarism Using Citations

This form is considered plagiarism as the text is referenced and reproduced with only superficial changes. It is still considered plagiarism as sections of this text should have been enclosed in quotation marks or footnotes should have been used (Hexham, 1999). The following example is taken from Hexham (1999):

- **Original Passage**

“But Hertzog recognized the danger and stood up for the rights of the Afrikaner. Only the National Party offered a Christian solution to South Africa's racial problems. The politics of the nationalists, were in the view of *Het Westen*, unquestionably Christian. The Afrikaner People were a Christian people, therefore their politics must of necessity be Christian.”<sup>1</sup>

<sup>1</sup> Irving Hexham, *The Irony of Apartheid* (Lewiston: Edwin Mellen, 1981), p. 185.

- **Plagiarised Passage**

“Professor Hexham brilliantly observes that Hertzog recognized the danger and stood up for the rights of the Afrikaner. Only the National Party offered a Christian solution to South Africa's racial problems. The politics of the nationalists, were in the view of *Het Westen*, unquestionably Christian. The Afrikaner People were a Christian people, therefore their politics must of necessity be Christian.”

- **Correct Passage**

“Professor Hexham observes, “Hertzog recognized the danger and stood up for the rights of the Afrikaner. Only the National Party offered a Christian solution to South Africa's racial problems. The politics of the nationalists, were in the view of *Het Westen*, unquestionably Christian. The Afrikaner People were a Christian people, therefore their politics must of necessity be Christian.”<sup>1</sup>

<sup>1</sup> Irving Hexham, *The Irony of Apartheid* (Lewiston: Edwin Mellen, 1981), p. 185.

Hexham (1999) notes that the use of exaggerated wording such as “brilliant” in the plagiarised paragraph can be used as an indicator that plagiarism is about to follow. He states that the use of such words is simply done to mask the culprit’s true intent. The other issue surrounding the plagiarised paragraph is that of quoting. The plagiarised version, if too similar to the original, should be enclosed in quotation marks.

Another form of plagiarism, where words are copied from a source and simply switched around, deleted, substituted, the verbatim text not enclosed in quotation marks and not referenced accordingly, is known as patch writing (Roig cites Howard, 1999), or paraphragiarism (Roig cites Levin & Marshall, 1993).

### 2.6.2.3 Patch Writing and Paraphrasing

According to the Simon Fraser University (SFU, 2007), patch writing occurs when text is utilised from another author’s source and not changed and re-interpreted enough, resulting in the new text being too similar to that of the original authors. Patch writing involves the deletion, switching and substitution of words as well as even leaving the grammatical style very close to the originals. Unlike patch writing, paraphrasing is not necessarily considered plagiarism. According to Hexham (1999), to ensure good paraphrasing, it must not make up the bulk of the written work but strive to get the same idea across using different wording. When paraphrasing, it is still essential to reference. The following examples taken from the SFU (2007) help to further illustrate patch writing and paraphrasing:

- **Original Passage**

“Where **mainstream sports** typically **refrain** from displaying **unapologetically violent acts**, professional wrestling dives in head first. **A large** portion of **wrestling’s** cultural **appeal is generated** by the psychological arousal/excitement provided by witnessing highly **aggressive and violent** forms of physical **interaction in this** sphere. Wrestling takes that which is pushed **behind the scenes of social life** and places it in the centre ring (Atkinson, 2002, pp. 62-63).”



- **Patch written version**

“**Mainstream sports refrain from** showing unremorseful **violent acts** while **professional wrestling** unapologetically revels in the same type of violence. **A large part of wrestling’s appeal is generated by the very aggressive and violent** interaction in this sport. While such violence is usually **behind the scenes of social life**, it is the centre of wrestling’s existence (Atkinson, 2002, pp. 62-63).”

- **Properly Paraphrased Version**

“Most sports do not encourage blatant acts of violence while professional wrestling embraces the same behaviour. Wrestling appeals to audiences because people enjoy watching aggressive and violent acts in the ring. What is normally not condoned in social life is made acceptable in wrestling (Atkinson, 2002, pp. 62-63).”

The words in bold in the patch written version reflect those which are too close to the pieces of text in the original passage. The effective use of paraphrasing is what is required to ensure the problem of patch writing does not occur. This being said, in order to paraphrase well one needs to possess a relatively good understanding of the material being used, thereby reducing the amount of exact and similar words utilised in the new text (SFU, 2007). Roig (2006) reiterates this point by stating that “In order to make substantial modifications to the original text that result in a proper paraphrase, the author must have a thorough understanding of the ideas and terminology being used”. The SFU (2007) also notes that paraphrasing is not a simple task. Roig (2006) states that whether information is paraphrased or even taken in quotation marks, the authors of the information must be acknowledged. He (2006) suggests that when paraphrasing one must ensure that one’s own words and writing styles are used to represent the ideas in the original work as well as being properly referenced.

Barnbaum (2002) identifies the following as the most common forms of plagiarism.

### **i. Copy and Paste Plagiarism**

This occurs when a whole sentence or significant phrase is taken whole from a source. Almost every word-processing software package allows text to be cut and paste from one document to another. This is the easiest form to detect.

### **ii. Word Switch Plagiarism**

In this instance a sentence is taken from a source and a few words contained within are switched around. This can be more of a challenge to detect, more so when the majority of the words are switched.

### **iii. Metaphor Plagiarism**

A metaphor is defined as the application of a name or descriptive term or phrase to an object or action where it is not literally applicable (Branford, 1991). This being said, it can be understood that metaphors give additional meaning to words and are an important part of the author's style.

### **iv. Idea Plagiarism**

When an author expresses a creative idea or a solution to a problem, the author must be given credit for it. Barnbaum (2002) notes that idea plagiarism is one of the most difficult forms of plagiarism for students to be able to distinguish. One must be able to distinguish from common knowledge, which does not need to be credited as it is taken for granted by everyone.

### **v. Reasoning Style/Organisational Plagiarism**

When one follows a source article's structure sentence by sentence, it does not matter if the sentences are not the same as it is still is plagiarism. A particular article may have a well-thought up structure and sequence of reasoning. Even if the words and concepts in the said article are written differently, it is still considered plagiarism if one follows the exact same style as the article in question.

## **2.6.3 Collusion**

According to the SFU (2007), collusion involves any form of unauthorised assistance that may be given when writing certain articles. An example would be a

lecturer who is paid by a student to complete a thesis for the student and which is then passed on to the university as the student's own work. Collusion is regarded as a form of plagiarism (SFU, 2007). It is also noted that when one is not sure as to what type or extent of assistance is allowed, one must get further clarity from the necessary powers that be (SFU, 2007).

#### **2.6.4 Self-Plagiarism**

Self-plagiarism is not among the more well-known forms of plagiarism and occurs when the author attempts to deceive the reader (Hexham, 1999). The concept of self-plagiarism may be difficult to conceptualise, especially if one looks at it from the point of view, "can I steal from myself"? However, Hexham (1999) also notes that according to law, it is possible to steal from oneself e.g. insurance fraud. Thus it is also possible to plagiarise from oneself, thereby deceiving the reader. Roig (2006) states that self-plagiarism occurs when the writer attempts to deceive readers by effecting minor changes to a previously written document and then passing it on as a new publication. These views are also expressed by Davidson et al. (2003), who state that papers that which have already been published must not be resubmitted, and only in very special circumstances can they be allowed. Roig (2006) notes that this type of plagiarism is a form of deception and that the concept of self-plagiarism can be broken down into four categories: redundant and duplicate publications, academic self-plagiarism and Salami Slicing.

##### **2.6.4.1 Redundant and Duplicate Publications**

According to Roig (2006), the current situation for the rewarding of academics places pressure on them to publish a large number of articles in order to secure bonuses and promotions. This scenario has a serious flaw in that duplicate and redundant publications are becoming more of a problem. Duplicate publications describe the publishing of the same data in more than one journal without bringing to light the fact that the information exists in other journals. There are some situations where duplicate publications are allowed, where the journals involved would collectively agree to the arrangement. The practice of redundant publications occurs when the same paper is published with a few superficial changes, i.e. the same data are published within a different context. Roig (2006) states that when an author wishes to submit an article containing data already available in another journal or

repository for instance, the author must make these facts known to the editors and readers.

#### **2.6.4.2 Academic Self-Plagiarism**

This instance of plagiarism occurs when students hand in the same paper or a substantial amount of it to different lecturers teaching different subjects within an academic institution (Roig, 2006). It also occurs when academics submit the same or very similar papers for publication to two different publishers in order to gain extra credit.

#### **2.6.4.3 Salami Slicing (Data Fragmentation and Augmentation)**

This occurs when a large article is split into two or more smaller ones. The inherent issue here is that readers are led to believe that the data in the articles are formulated from different sources. Data augmentation is closely related to the concept of fragmentation; however, augmentation describes a situation where an article is published and then later on, the author collects additional data strengthening the article, but instead publishes the results as a new publication (Roig, 2006). A guideline to avoid this situation is that should any doubt exist in the writer's mind as to whether a paper to be submitted contains fragmented data, the author should attach the other papers that contain the fragmented data to the article (Roig, 2006).

#### **2.6.5 Effects of Plagiarism at Academic Institutions**

Plagiarism can lead to many problems for an academic institution; this section aims to highlight the more pertinent issues. The damage caused by plagiarism is not always what first meets the eye, as plagiarism has far reaching consequences. These range from undermining the standard of academic work to the commission of other crimes later on in a plagiariser's life. The following points created by Swarthmore College University [hereby refer to as Swarthmore College (2004a)], depicts some of the ill effects plagiarism can have on an academic institution.

##### **2.6.5.1 Plagiarism causes Grade Inflation**

Students who commit plagiarism tend to obtain higher grades than they would normally achieve. With plagiarism it is possible to get good grades, far better than the student would have achieved had he/she done the work on their own. If a number of

students plagiarise, the averages for that particular subject will usually appear higher than they should actually be.

#### **2.6.5.2 Plagiarism hinders the Proper Pitching of Assignments**

If plagiarism is common and overlooked, the departments could be convinced that the standards are low, given the resulting high grades. In response they will increase the difficulty of the academic work to make the averages drop to a lower level and to make the course work challenging for students.

#### **2.6.5.3 Plagiarism Cases at academic institutions are Huge Time Sinks**

This refers to the fact that plagiarism cases tried at universities are very time-consuming. Additionally, they also usually require a panel of staff present at the hearings and could incur costs.

#### **2.6.5.4 Encountering Plagiarism is depressing for Academic Institutions**

Incidents of plagiarism at any university will inevitably undermine the institution's prestige. For an academic institution, reputation is of utmost importance and is what it relies on to attract applicants and remain competitive.

#### **2.6.5.5 Plagiarisers undermine their own Education**

If students get away by using other peoples' work, then they themselves do not benefit from fully engaging the topics and/or assignments. Part of the learning process is researching the material oneself. Plagiarism destroys any sort of initiative by seemingly proposing a quick solution.

#### **2.6.5.6 Successful plagiarism encourages Lifelong Dishonesty**

Undergraduates that plagiarise are more prone to committing more serious and repetitive crimes later on in life (Swarthmore College, 2004a).

#### **2.6.5.7 Rampant plagiarism demoralises Honest Students**

Should dishonest students achieve better grades than those who work hard and get away with plagiarising, this would demoralise the hard working students who do not cheat, as well as diminish their faith in the institution's moral standards.

## **2.6.6 Techniques in the Deterrence of Plagiarism**

Rather than waiting for the crime to occur, one should strive to prevent it from taking place in the first instance. "Prevention is better than cure" is a common saying which is applicable in this context. The following points were noted by Swarthmore College (2004b):

### **2.6.6.1 Add Serious Anti-Plagiarism Warnings to the Syllabus**

Adding warnings on the repercussions of plagiarism will make students think twice about copying/plagiarising work. However, a study by O'Connor (2003) showed that a warning might not make such a difference. In the experiment, two teachers taking the same course gave half the class serious warnings about plagiarism, while the other half was given none. The results were run through plagiarism detection software and equal amounts of cheating were found.

### **2.6.6.2 Tell Students how the institution detects Plagiarism**

The most common characteristic of plagiarised work is the difference in the grammar style. Informing students that the faculty can detect such differences in grammar is in fact telling them that, the university is indeed looking for plagiarism (Swarthmore College, 2004b).

### **2.6.6.3 Make it known that the university uses Plagiarism-Detection Software**

The fact that the university uses such technology to catch plagiarisers is a serious deterrent for those deciding whether or not to plagiarise. Even "saying" you will use detection software has been shown to be more effective than telling students, "don't plagiarise." (Swarthmore College, 2004b).

### **2.6.6.4 Inform students plagiarisers will have to appear before a Judicial Committee**

Informing students that they could be expelled from the university should they be caught plagiarising, is a deterrent. Including rules such as this in the university rule book should be considered essential; no student should be able to claim ignorance of this fact.

## **2.6.7 Techniques to Detect Plagiarism**

Deterring potential offenders is one aspect of the plagiarism problem; the other issue revolves around those individuals who do not fear threats and the consequences of unethical actions. They may also feel the benefit clearly outweighs the consequences of getting caught. For these individuals, the university needs to enforce several different methods to root out offenders. These consist of software solutions, outsourcing and online solutions. Factors such as price, convenience and effectiveness will be taken into account. A framework for analysing these various packages will be created or located.

There are several methods that can be used to determine if plagiarism exists in an article. One can start by noticing changes in grammatical/writing styles; for example, the majority of the article has misspelt wording, yet the remainder is perfect. Then, those perfect paragraphs need to be looked at more closely (Hexham, 1999). This is an effective method as it bypasses the issues of word switching to an extent and helps narrow down the plagiarised portions. Another method to identify plagiarism is when a paraphrased passage contains the work of a major author yet the reference is that of another author merely *re-interpreting* the original work, but passed on as if it is the original work (Hexham, 1999). Hexham (1999) also notes that many plagiarisers are often given away by plagiarising the mistakes in the original author's works.

### **2.6.7.1 Software**

There are software packages available for purchase to combat the problem of plagiarism. The following discussion outlines some of the various software tools that can be used to detect plagiarism. It is important to note that these are only a few of the various programs available and that the aim of this section is to provide a brief understanding of the features, functions and capabilities of these programs. The programs presented in this study will not be compared or analysed in depth as this falls outside of the scope of its research. The features discussed are summarised from the respective websites.

### **2.6.7.1.1 CopyCatch Investigator**

The following facts and points are closely summarised from the CopyCatchGold (2009) program, which is referenced in the bibliography.

#### **i. Features (Taken verbatim from the CopyCatchGold web site)**

- GUI or automatic – it can be used through a graphical user interface or through command line instructions
- Multi-threaded, multi-processor capability – modern computers often have multiple processor cores built into one chip. These cores can be used in parallel to further enhance the processing speed
- Multi-platform - written in Java – this essentially means the program can be used on a variety of different operating systems

#### **ii. Function**

This package searches for similarities in sentences and documents. Here entire texts or multiple sets of texts are used as the search data. The user then sets the level of similarity to be detected. Similarity in this scenario refers to the level at which the system attempts to match the data entered. The higher the level of similarity, the less the number of results will be outputted.

#### **iii. The interface**

The CopyCatch user interface is simple to operate and contains a search tab, indexing tab, content words tab and a statistics tab.

#### **iv. The outputting of data**

When similar pieces of text are located they are displayed next to each other and are linked to the position in the respective document. Both documents can then be fully viewed next to each other.

#### **v. The report**

The report that is created contains levels of details that the user specifies. The minimum number of sentences in common can be specified, as well as the minimum



number of words that must match in a sentence and lastly, the level of sentence similarity can be set according to a percentage. All these levels are set by using a slider and moving it up or down.

#### **vi. Multilingual**

The program can be inputted with a list of function words, which can be obtained from the CopyCatch organisation. The program cannot look for similarities in documents with different languages.

#### **2.6.7.1.2 SafeAssign (MyDropBox)**

The following facts and points are closely summarised from the Knight and Minkoff (2009) - SafeAssign program, which is referenced in the bibliography.

The SafeAssign program is used to detect plagiarism in student papers. It also contains features to aid in educating students about the dangers of plagiarism.

#### **i. Usage**

- Students can submit their assignments to the SafeAssign server and have them automatically checked against Internet data for plagiarism. These assignments can then be forwarded straight to the lecturer with the corresponding report.
- The other method is for instructors to directly submit student assignments to SafeAssign for review.

#### **ii. How it works**

SafeAssign uses the following repositories to match for instances of plagiarism.

- The Internet
- ProQuest ABI/Inform database, which contains hundreds of published titles and over two million articles updated weekly
- Institutional document archives; this is the repository that contains papers submitted by users in their respective institutions

- Global Reference Database; this is a repository of papers that were uploaded by students from Blackboard client institutions to prevent plagiarism

#### **2.6.7.1.3 Eve2**

The following facts and points are closely summarised from the Canexus (No Date) web site, which is referenced in the bibliography.

Eve2 is an anti-plagiarism program that links to the Internet to search for instances of plagiarism. Documents are inputted into the program in plain text format, and then Eve2 searches the Internet according to the user's selection of either "quick", "medium", or "extra strength". Once this process is completed, the program outputs a results file containing the links to the various site/s that have been plagiarised.

##### **i. Using Eve2**

Eve2 accepts any of the following formats: Plain Text, Microsoft Word, and Word Perfect. It is important to note that Eve2 only shows the plagiarised web links for plain text files (.txt); therefore one can simply run all the file formats and once the list of files that plagiarism has been detected in is outputted, one can convert those into plain text format and re-run the detection in order to get the URLs of the sites that have been plagiarised from.

The program can be set to stop when a certain percentage of plagiarism has been detected in a document, thereby saving time. While the program searches the Internet, a status bar will be depicted indicating the total progress on the particular file. Should plagiarism be detected, the program will output a results page showing each document and the corresponding percent of plagiarism. These individual documents can then be clicked on to show greater detail.

#### **2.6.7.1.4 Glatt Plagiarism Screening Program**

The following facts and points are closely summarised from the Glatt Plagiarism Screening Program (No Date), which is referenced in the bibliography.

This program was created to allow users to search for plagiarism according to writing style. Each person has a unique writing style. The program works by

removing every fifth word in a student's paper and then asking the student to supply the missing words. The plagiarism probability is calculated on the number of correct responses and the amount of time intervening as well as other factors. This package took approximately ten years to develop and has been extensively tested. The developers state that not a single student has ever been falsely accused of plagiarism. This program is especially useful in situations where the original material cannot be located.

#### **2.6.7.1.5 JPlag**

The following facts and points are closely summarised from the Jplag (No Date) program, which is referenced in the bibliography.

Jplag is an anti-plagiarism program centred towards the programming environment. It compares both programming syntax and structure, not just text. This makes it far more robust and useful. Jplag supports Java, C#, C, C++, Scheme and natural language text. Jplag has been used in several intellectual property cases and by expert witnesses. It does not link to the Internet to perform content searches. The package is free for use; however, one must obtain an account before usage.

#### **2.6.7.1.6 PlagiarismDetect**

The following facts and points are closely summarised from the PlagiarismDetect (2009) program, which is referenced in the bibliography.

##### **i. Features**

- Uses an advanced algorithm, with multi-layered detection capabilities

PlagiarismDetect contains three advanced detection algorithms. After all three algorithms are run, the program outputs a report detailing the levels of plagiarism found.

- Directly upload commonly used document formats

This program accepts the uploading of the following file formats for analysis: .txt (text file), .doc (MS Word), .docx (MS Word), .rtf (Rich Text),

.xls (MS Excel), .xlsx (MS Excel), .ppt(MS Powerpoint) and pptx (MS Powerpoint)

- Batch upload option

This function was created specifically for academics and simply put, allows for the uploading of up to twenty documents at a time. The program can then scan all the documents in the batch and output a report to the users' email address. This is a very handy function to have in situations where one wishes to assess an entire group of assignments at once.

- Comparison of a document with a database of documents

This function allows for the user to check one document against a database of texts that could reside on a local computer or data store.

- Searching against a specific URL

This function allows for searches to be performed against a specific URL to determine if anything from the document was plagiarised.

- Microsoft Office 2007 plug-in for detecting plagiarism

This feature allows for plagiarism detection to be run directly from Office 2007, as it is a plug-in that installs into the MS Office toolbar.

- Avoiding Self Plagiarism

This program will allow the user to search for instances of self-plagiarism and avoid it.

#### **2.6.7.1.7 Pl@giarism**

The following facts and points are closely summarised from the Pl@giarism (2008) program, which is referenced in the bibliography.

### **i. The program in practice**

The Pl@giarism program uses a table system to display resemblance percentages of documents. Matching text will be shown in blue.

### **ii. How the program works**

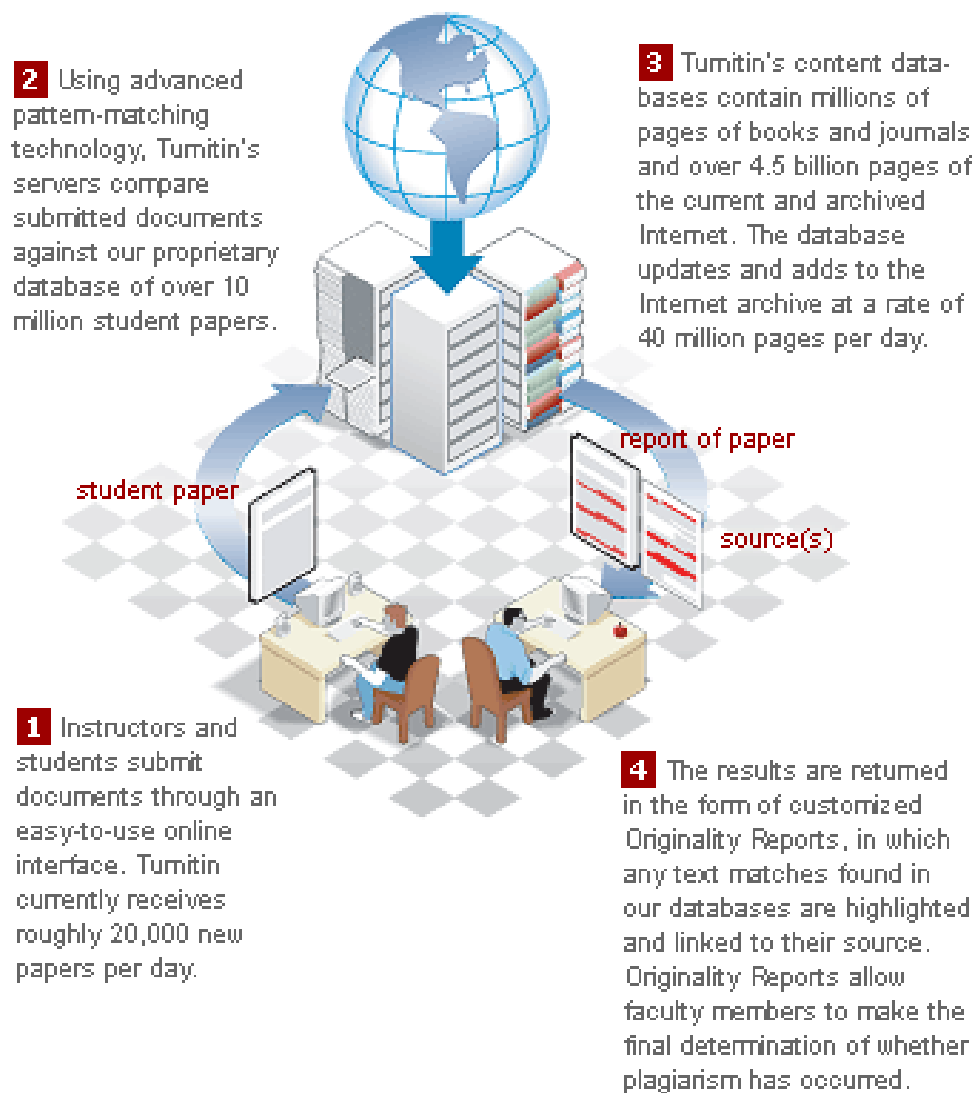
- This tool is capable of running searches on multiple documents in one folder, a handy feature as one can simply place all material within that folder and perform a search on all the documents.
- By scanning the Internet for documents that match a phrase from a selected document. This feature is insufficient for scanning the whole document for plagiarism, and it will only search with known search engines.
- The program links to the Internet to run searches for sections of text in the documents. Only known search engines will be used to perform this task. Documents can be compared to a list of texts and phrases can be excluded. The program makes use of filters to remove documents the user does not want to see.

### **iii. Information**

- The program has been created in Microsoft Visual Basic 6.0 and Microsoft Word is required to utilise the program. It is noted that the program does take a fairly long time to detect plagiarism due to its meticulous processes.

## **2.6.7.2 Outsourcing and Online Solutions**

There are companies that offer their anti-plagiarism services e.g. *Turnitin.com*. This particular company has a database of four and a half billion web pages updated at a rate of forty million pages per day (Turnitin.com, 2005b). Its database also consists of student papers and published works.



**Figure 2.20: Turnitin's Plagiarism Prevention System, adopted from: Turnitin.com (2004)**

Figure 2.20 depicts Turnitin's plagiarism prevention system. To summarise, documents are submitted and then the servers compare the data to that which they have collected in their databases. Once this process is completed, the results are returned with text matches highlighted and linked to the source document. Turnitin uses a colour coded system to present results of submitted work. According to O'Connor (2003), red indicates that 75% of the work has been detected as being possible instances of plagiarism, while yellow indicates that 50% was and so on. His study further found that when one thousand nine hundred and twenty five essays across a range of twenty subjects in six Victorian universities were submitted to Turnitin, the overall result was that almost 14% of all the essays contained

unacceptable levels of plagiarism. The results also proved that one student had copied 90% of the essay in several chunks directly from the Internet, only writing the conclusion paragraph him/herself. The study concluded that most universities will have a sizable amount of plagiarism occurring.

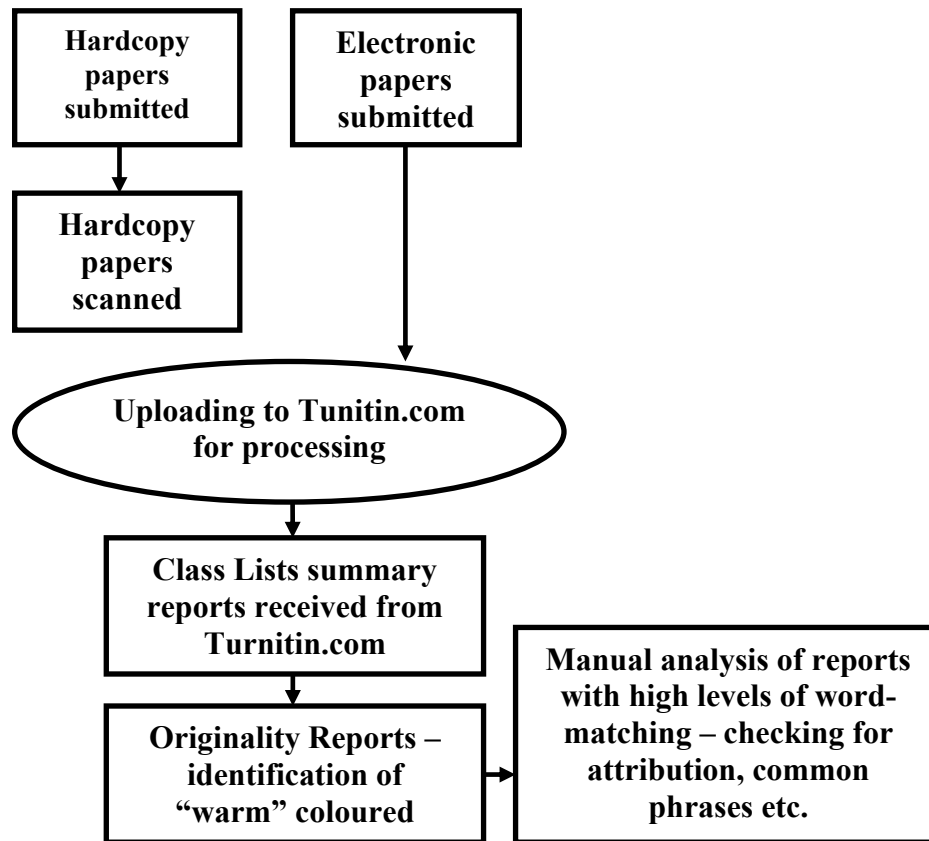


Figure 2.21: Caval project process, from: O'Connor - Cheating and electronic plagiarism [2003]

Figure 2.21 indicates the process used to analyse the level of plagiarism in the one thousand nine hundred and twenty five essays. A recent study by Turnitin.com (a) (2005, cited in Rutgers University/Centre for Academic Integrity Study, 2003) showed that 38% of students admitted to cut-and-paste Internet plagiarism in the previous year. Another study by Turnitin showed that 80% of college-bound students admit to cheating on schoolwork, yet 95% of them never get caught (Turnitin.com cites Who's Who Among American High School Students 2005).

#### 2.6.7.2.1 Google as an Anti-Plagiarism Tool

According to Weeks (2006), Google can be a potent tool in the fight against plagiarism. The writer notes a 75% success rate when using Google to detect

plagiarism. Here random pieces of text are taken usually using names or unusual phrases of around six words. This shows that the search engine Google is a useful and effective tool in this regard.

#### **2.6.7.2.2 CopyScape**

Copyscape is a free online service to determine if there are copies of one's content on the Internet (CopyScape, 2009). The program is very simplistic and the website does not give much detail on its functioning.

#### **2.6.7.2.3 Article Checker**

This online tool can be located at Article Checker (2008).

This program is very similar in nature to the CopyScape program described in the previous point. Article Checker also features a text search, where text is simply pasted into the text search box. The program is very simplistic in nature.

#### **2.6.7.2.4 Dupli Checker**

The following facts and points are closely summarised from the Dupli Checker (2008) program, which is referenced in the bibliography.

This is yet another simplistic online anti-plagiarism tool. The following points are the instructions for using the tool.

- i.** Type few phrases or add your article in the box on the web page
- ii.** Choose either quote or without quote and the search engine, then press the search button
- iii.** A new search page will show results after searching each sentence against the websites pages that are already indexed

#### **2.6.8 Prosecution of Offenders**

The question of the extent one punishes a plagiariser is a rather difficult one, mainly due to the fact that plagiarism is not a simple, black and white matter, but rather has many shades of grey. It is important to firstly check that the issues with the material in question have not been made by the printers or editors of the article, as one



cannot hold the author liable if he/she did not write the plagiarised words (Hexham, 1999). It is also very important to note where the alleged text is from, as anything that could be classified as being part of the body of common knowledge or lies in the open domain cannot be classed as copyrighted work. Another important factor to consider is that of intent and whether it is academic or other writing (Hexham, 1999). Is the author being accused of using a description from another article or using an innovative idea?

### **2.6.9 Plagiarism and Cultural Relationships**

The question of whether plagiarism has any cultural relationships is a difficult one. Our society is made up of many different people who come from very different backgrounds. These different cultures have very different rules driven by religion. Pecorari (2003) notes that students who come from background cultures and beliefs where emphasis is placed on collective effort, may not fully understand the concept of plagiarism. To these individuals, copying from another's work without giving acknowledgement to the author may seem irrelevant. The author also notes that there is much uncertainty as to why students from these same backgrounds, who are informed and educated as to what plagiarism is, still commit the offence.

### **2.7 Stylistics and Linguistics**

These two terms go hand-in-hand as stylistics encompasses the scrutiny of text in order to determine patterns or deviations in writing styles. AskOxford.com (2007b) defines stylistics as “the study of literary styles of particular genres or writers”. The term “linguistics” is also defined as “the scientific study of language and its structure” (AskOxford.com, 2007c). According to Wynne (2005), stylistics may be defined as the “study of the language of literature”; he also goes on to say that this study of language makes use of various linguistic tools in order to analyse the texts. Missikova (2003) notes that many different stylistic approaches exist due to current literary criticism. Another important fact is that the goal of stylistic analysis is not simply to determine the features of text, but rather to show the word's functional significance for the interpretation of the texts (Missikova, 2003). It is now clearer that stylistics and linguistics are very closely intertwined. Wynne (2005) also mentions that stylistics and linguistics are increasingly converging and will overlap as empirical work in linguistics increasingly makes use of stylistic methods. The reason for

barriers existing between linguistics and stylistics is because people lack the skill necessary to use computers for stylistic analysis (Wynne, 2005). Some of the problems encountered with the use of computers in this field are that the user will have to set constraints in the software, often done by programming the instructions. A person without relatively advanced computer skills will find this difficult to do. Wynne (2005) also states that even though texts do appear in electronic format, they often vary in layout, quality and file formats, among other attributes. This is where stylistic tools come into play, as some of these packages can bring a document down to its raw format. This allows for various types and formats of documents to be analysed together.

Stylistics, however, has many variations and methods of usage. According to Pecorari (2003), a study was done based on a selection of seventeen masters' dissertations and PhD theses. To summarise, the study had shown that all seventeen documents contained elements of plagiarism. The degree of plagiarism varied greatly, but in the worst case one document was found with 95% of its contents being plagiarised. The average plagiarism amongst the documents was around 30-45%. These results are very serious and highlight an urgent need for academic institutions to educate students on how to best avoid plagiarism and the consequences thereof if one is found guilty of the offence.

### **2.7.1 Corpus-Based Text Analysis**

Kilgarriff (1997) states that a corpus is a collection of texts that could range from a few words to thousands. These samples can come from both speech and writing. According to McEnery and Wilson (No Date), a corpus is a collection of more than one text. Corpus usage is imperative due to the fact that computers are now able to analyse corpora in a way that was not possible in the pre-computational era (Hong Kong Polytechnic University, [No Date]). This allows for much more advanced stylistic/pattern methods to be utilised in order to determine authorship and other factors of the texts. The term corpus annotation is one that is not widely used and simply means the adding of value to a body of text. This feature is explained in more detail below.

### 2.7.2 Corpus Annotation

Corpus annotation simply means adding additional interpretive material or tags to the corpus. Tags, according to Wynne (2005), are intended to add additional value to the corpus by facilitating for categorisation of stylistic features. Annotation is particularly useful when distinguishing between words which have the same spelling but different meanings, depending on how the word is utilised in a sentence (e.g. the word “present” could mean a “gift”, which is a noun or “give someone a present” which is a verb or “he was present” which now becomes a verb). Tagging text helps distinguish these various forms of meaning and gives “added value” to the corpus (Leech, 2004). These tags would be described as follows:

present\_NN1 (singular common noun)

present\_VVB (base form of a lexical verb)

present\_JJ (general adjective)

Leech (2004) identifies the following different kinds of annotation:

- Phonetic Annotation

This refers to the way in which a particular word in the corpus sounds. Here information such as stress and pauses are noted, basically breaking down the sentence into units such as phrases and clauses.

- Semantic Annotation

Here information on the different meanings of words is ascertained (e.g. “cricket” refers to the sport as well as the insect).

- Discourse Annotation

Used to describe words that refer to a grouping (e.g. “them” and the horses are running bring “them” around).

- Stylistic Annotation

Information about speech and thought presentation.

- Lexical Annotation

This encompasses the analysis of base forms of words such as “Lie” and “Lying”.

### **2.7.3 Function Word Analysis**

Many different techniques exist for determining authorship in written texts. Function word analysis is one of these techniques. Argamon and Levitan (2005) cite Mosteller and Wallace (No date), stating that function word analysis was created around forty years ago and that function words are a collection of a small number of the most frequent words in language. These words could then be utilised to determine authorship. Up-to-date, more advanced algorithms have been created to utilise these function words for stylistic analysis. However, while these improvements have only made the output more reliable, they have only barely been able to improve on the usefulness of the function word analysis. According to Argamon and Levitan (2005), the underlying effectiveness of function words is attributed to their high frequency of use in language that authors have very limited conscious control over. This essentially means that each author/writer will have his or her own unique style of writing based on the usage of these function words. The analysis of these function words can also be extended to the analysis of individual function word usage. This is imperative as the usage of individual function words will vary among different authors, once again allowing for a stylistic pattern to be created (Argamon and Levitan, 2005). Another method was developed by Hoover, who created a system using frequent word occurrences or collections (e.g. the word “and” appearing five words away from the word “is”). A word collection is defined as “a certain pair of words occurring within a given threshold distance of each other” (Argamon and Levitan, 2005). This system works in such a way that the chosen pairs or combinations of words are determined over the entire analysed text; it is also noted that these stylistic systems become more effective the longer the corpus is. However, research has also shown that if Hoover’s system is used on a corpus of ten thousand words or less, the reliability of this system is greatly reduced and therefore will not be able to yield very reliable results. Research has also revealed that function words are superior to the usage of word collections (Argamon and Levitan, 2005).

AskOxford.com (2007a) explains that just ten different function words (the, be, to, of, and, a, in, that, have, and I) account for approximately twenty five percent of the total words (one billion) used in their corpus.

Vocabulary size (no. lemmas)	% of content in OEC	Example lemmas
10	25%	the, of, and, to, that, have
100	50%	from, because, go, me, our, well, way
1000	75%	girl, win, decide, huge, difficult, series
7000	90%	tackle, peak, crude, purely, dude, modest
50,000	95%	saboteur, autocracy, calyx, conformist
>1,000,000	99%	laggardly, endobenthic, pomological

**Table 2.1: Function Word Usage in the Oxford English Corpus, from: AskOxford.com (2007a)**

As indicated on the above table, it is interesting to note that as the number of function words or lemmas increases, the percentage of the corpus that is used up by the function words also increases. The table also shows that as a language user's vocabulary size increases, so does the complexity of the words that are used. A lower vocabulary size will consist of higher frequency words (more commonly used). According to AskOxford.com (2007a), the 100 most common words are depicted in the table below.

1 the	26 they	51 when	76 come
2 be	27 we	52 make	77 its
3 to	28 say	53 can	78 over
4 of	29 her	54 like	79 think
5 and	30 she	55 time	80 also
6 a	31 or	56 no	81 back
7 in	32 an	57 just	82 after
8 that	33 will	58 him	83 use
9 have	34 my	59 know	84 two
10 I	35 one	60 take	85 how
11 it	36 all	61 people	86 our
12 for	37 would	62 into	87 work
13 not	38 there	63 year	88 first
14 on	39 their	64 your	89 well
15 with	40 what	65 good	90 way
16 he	41 so	66 some	91 even
17 as	42 up	67 could	92 new
18 you	43 out	68 them	93 want
19 do	44 if	69 see	94 because
20 at	45 about	70 other	95 any
21 this	46 who	71 than	96 these
22 but	47 get	72 then	97 give
23 his	48 which	73 now	98 day
24 by	49 go	74 look	99 most
25 from	50 me	75 only	100 us

**Table 2.2: The 100 most common Function Words in Oxford English Corpus, from: AskOxford.com (2007a)**

Table 2.2 above gives an accurate indication of the most commonly used words in the Oxford English corpus, with the word “the” being the most common and the word “us” being the least commonly used word out of the list of 100 most commonly used words. These function words will play a pivotal role in the analysis of corpuses in Chapter 3. The usage of function words depicted in the table above differ from author to author; it is this fact that will enable the researcher to draw comparisons and accurate statistics based on an individual’s written work. These function words can further be broken down into groups such as nouns, verbs and adjectives. The table below depicts this breakdown.

Nouns		Verbs		Adjectives	
1	time	1	be	1	good
2	person	2	have	2	new
3	year	3	do	3	first
4	way	4	say	4	last
5	day	5	get	5	long
6	thing	6	make	6	great
7	man	7	go	7	little
8	world	8	know	8	own
9	life	9	take	9	other
10	hand	10	see	10	old
11	part	11	come	11	right
12	child	12	think	12	big
13	eye	13	look	13	high
14	woman	14	want	14	different
15	place	15	give	15	small
16	work	16	use	16	large
17	week	17	find	17	next
18	case	18	tell	18	early
19	point	19	ask	19	young
20	government	20	work	20	important
21	company	21	seem	21	few
22	number	22	feel	22	public
23	group	23	try	23	bad
24	problem	24	leave	24	same
25	fact	25	call	25	able

**Table 2.3: Top 25 Function Words Classified as Nouns, Verbs and Adjectives in Oxford English Corpus, from: AskOxford.com (2007a)**

Table 2.3 shows the top twenty-five most used function words that are nouns, verbs and adjectives. A noun is defined as “a word that refers to a person, place, thing, event, substance or quality”, a verb is defined as “a word or phrase that describes an action, condition or experience” and an adjective is defined as “a word that describes a noun or pronoun” (Cambridge, 2007). These three categories can further aid with the determination of authorship as more statistical analysis can now be performed on

a particular author's work. The rarer the word, the lower is its frequency of usage and *vice versa*. It is but one of the techniques that can be used for determining authorship. It is important to note that in order to ensure accurate results, these techniques must be used together. There are many other stylistic analysis techniques in existence, which, however, fall outside the scope of this study.

## **2.8 Models and Tools used in e-Forensics**

A number of different forensic models exist e.g. figure 2.26 CTOSE Reference Process Model; however, there are only a few when it comes to forensic computing and even fewer that have been specifically created with plagiarism and stylistics in mind. There are a number of different types of forensics in existence e.g. forensic pathology and forensic ballistics. Each of these types will have their own processes and models. They will also have their own unique methods and approaches when implemented. The importance of a sound forensic computing plan cannot be understated, as the plan needs to be implemented as soon as a problem is detected. It also serves to ensure the forensic process complies with legal requirements and technicalities that could invalidate the entire case in a court of law should the investigation be performed incorrectly.

## **2.9 Legal Implications**

In order to conduct a successful forensic computing analysis, one must abide by the rules and regulations of the law. This is by no means an easy task, as the law is very complex and intricate. It would be a waste of time and money to conduct an investigation only to find that the manner in which the evidence was acquired rendered it invalid in a court of law. These legal aspects play a pivotal role in the prosecution of plagiarisers as well as any weighing heavily on the particular methods that are chosen when going about an investigation of the offence. These legal rules and laws will be covered in a later study as they fall out of the scope of this one.

## **2.10 The e-Forensic Life Cycle**

The e-forensic life cycle relates to the steps and procedures necessary to prosecute an electronic crime, and to encompass these steps one needs to utilise a forensic framework. There are a number of forensic frameworks in existence; however, the bulk of these relate to different branches of forensics. Very few

frameworks are in existence with regard to the prosecution of plagiarism or crimes of this nature.

### **2.10.1 Understanding the Cyber Forensic Investigation Process**

In order to go about conducting a forensic investigation of this nature, one must understand the pertinent points surrounding a cyber forensic investigation. Carrier and Spafford (2004), explain the forensic process to be a combination of several factors. Firstly, digital data are represented in a numerical form, also known as binary which is a collection of 0's and 1's. The digital data have to be stored somewhere on a device e.g. hard disk drives and flash drives. Digital objects are described as a collection of digital data e.g. a file or a process. Each digital object has different properties that describe it on the storage medium. It is these characteristics that can be used to track changes in these files. Carrier and Spafford (2004) state that a digital event can be described as an occurrence that changes the state of a digital object. Should an object have the ability to create an event, then it is known as a cause and should this event change another object's state, then this changed object is known as evidence of an event. An incident or digital incident is best described as an event or events that violate policy, not necessarily the law. A crime, on the other hand, is an event or group of events that violates the law, and an investigation is described as a process that creates and tests theories about events that have occurred (Carrier and Spafford, 2004). Carrier and Spafford also define physical evidence of an incident as a "physical object that contains reliable information that supports or refutes a hypothesis about the incident". They add that digital evidence of an incident is defined as "any digital data that contain reliable information that supports or refutes a hypothesis about the incident". They also noted that to link the two aspects, a cause and effect relationship exists, as the physical object contains data that pertains to the incident. In this respect, a hard disk drive is the physical evidence and the sectors on it are the digital evidence. It is important to note that a cyber forensic investigation does not simply involve the copying of digital data as this has to be acquired using stringent rules and methods to ensure that the validity of the data remains unchanged. Computer forensics encompasses the in-depth analysis of data in cyberspace.



### **2.10.2 MD5 and CRC Authentication**

Bunting and Wei (2006) note that two concepts, the MD5 and CRC values, can be used to determine whether or not the original seized electronic data remain unchanged during the course of the investigation. It is imperative that the original data remain unchanged as a number of factors could have an influence on the seized data during the investigation phase e.g. viruses could modify files and thereby delete material imperative to the case. Should the original seized data become modified in any way, a court of law would deem the evidence inadmissible. Forensic tools are therefore used to calculate these MD5 and CRC values. Bunting and Wei (2006) add that the probability of two files having the same MD5 value is one in approximately three hundred and forty billion, billion, billion, billion, and the probability of two files having the same CRC value is one in 4,294,967,296. These two concepts are what give the investigation the integrity that is needed during the information-acquiring process. Bunting and Wei (2006) also add that calculating MD5 values require much more computer processing time as opposed to calculating CRC values.

ISACA (2004) notes that in order to go about a computer forensic investigation, one must acquire an assignment mandate. This mandate is a written document giving appropriate authority to the investigator to go about the investigation. ISACA (2004) states that the mandate should contain the following details:

- Responsibility, authority and limitations of the investigation
- Ensure the independence of the IS auditor
- State that the investigator is acting within lawful authority
- Specify the scope and responsibilities of external experts who may be called in to assist with the investigation

The issue of independence is thus a very pertinent factor when dealing with forensic investigations of any type.

### **2.11 Forensic Frameworks**

In order to successfully go about a cyber forensics investigation, one needs to implement an appropriate framework. Wordnet (2006) defines a framework as “a hypothetical description of a complex entity or process”. This definition means that a framework is a tool that can be created to simplify a complex set of processes. It adds

a sequential system to the existing process and can show the flow of data between the various entities or role players. For the purposes of this study, a framework will allow for a complex set of forensic steps, processes and procedures to be run, in correct sequence, so as to output an acceptable result. The need for a solid forensic framework is of paramount importance. Zatyko (2007) depicts a brief breakdown of forensic investigations as being encompassed in eight steps. These are the search authority, chain of custody, imaging/hashing functions, validated tools, analysis, quality assurance, reporting and possible expert presentation.

### **2.11.1 The Processes and Factors involved in a Digital Forensic Investigation**

ISACA (2004) states that the primary goal of computer forensics is to discover the truth. It adds that computer forensics can aid an organisation in protecting its assets and provide a better understanding of attackers and attacks.

#### **i. Five characteristics of good forensics guidelines, adapted from ISACA (2004):**

- To reiterate the importance of responding immediately to a breach, thereby reducing the risk of evidence being tampered with and being lost
- To acquire and preserve data as close to the incident as possible
- To preserve the evidence in a state that will ensure its admissibility in a court of law
- To minimise the disruption to the daily functioning of the organisation when the investigation is in progress
- To identify an attacker and acquire proof

#### **ii. Independence:**

The term “independence” in the realm of forensics refers to the issue of conflict of interest. It is of utmost importance that no conflicts of interest exist between the investigator and/or the investigation team and the client or person being investigated. A conflict of interest could lead to bias in an investigation and render the process invalid in a court of law (ISACA, 2004).

### **iii. Digital Signatures:**

Digital signatures can be used to prove authenticity of a document, provided it had been signed using this method. Documents containing digital signatures are admissible in a court of law, so long as the key is authentic and it can be proved that should the document be modified by anyone other than the key holder, the key will be compromised, thereby revealing an unauthorised change. Another important factor to take into account is that should a person's key get mislaid or stolen, any documents signed within that period of time will be taken to be compromised and will not hold up as evidence in a court of law (ISACA, 2004).

#### **2.11.2 Elements when Planning a Computer Forensic Investigation**

Before going about an electronic forensic investigation, it is important to plan for the anticipated situation. Beginning the investigation on a solid grounding could have huge financial benefits, not to mention save time. The following factors embody some of the points to consider.

##### **2.11.2.1 Data Protection**

This relates to the protection of relevant data on the electronic devices, as well as the physical requirement for the protection of these devices. ISACA (2004) notes that it is important for the relevant parties to be informed regarding the requirements for certain evidence and to inform them not to destroy anything that may be pertinent to the case.

##### **2.11.2.2 Data Acquisition**

Data acquisition is primarily concerned with the capturing of data into a secure location; this could be from a variety of electronic devices and recorded statements from persons related to the case. ISACA (2004) notes that when this process is taking place, it is of utmost importance that the media is write-protected and virus-free. In this phase, the "taking down" (See section 2.13.1) of computer equipment is important.

##### **2.11.2.3 Imaging**

Imaging is the 1:1 copying of data; for forensic purposes this means exact duplicate. Bunting and Wei (2006) describe an image file to be a legal copy of the

original evidence, which can be used by the investigator for data analysis purposes. Imaging has the added benefit of protecting the original data from being modified.

Carrier and Spafford (2004) outline a framework for crime scene investigations (figure 2.22), and then adapt the crime scene framework so that it would aid in the solving of digital crimes. This framework consists of three phases and is known as the “digital crime scene investigation framework”. This research shall in a similar manner take this process one step further, by devising a conceptual framework for the digital forensic auditing of documents.

#### **2.11.2.4 Data Extraction**

Once the image or duplicate of the original data is made, the process of data extraction can begin. This process comprises of searching for relevant evidence and data objects within the image file. This data recovery process could be of damaged, destroyed or corrupted data (ISACA, 2004). Legally sound programs must be used to facilitate this search and extraction process.

#### **2.11.2.5 Data Standardisation for use by Non-Technical Persons**

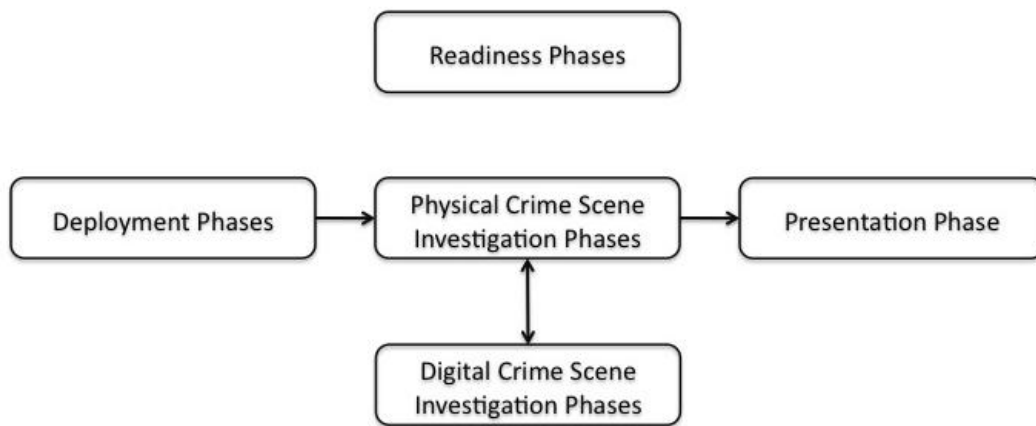
This process involves the organisation of the extracted media into a format easily understood by investigators and other role players. This process could involve creating searchable indexed databases of the compiled evidence using software programs such as DtSearch.

#### **2.11.2.6 Reporting on the Findings**

According to ISACA (2004), a final report that is to be presented to the court should contain the following details: scope, objectives, nature, timing and extent of the investigation. The findings, conclusions, and recommendations as well as investigators’ qualifications should also be clearly presented. The presentation of the evidence is a factor not to be taken lightly; ISACA (2004) states that due to the fact that electronic evidence can be found on a number of different media types, industry best practices must be employed and legally sound tools used to perform the extraction and analysis of data.

### **2.12 Crime Scene Investigation Framework**

The phases of the crime scene investigation framework follow:



**Figure 2.22: Major Categories of Phases in the Crime Scene Investigation Framework, adopted from: Carrier and Spafford (2004)**

Figure 2.22 above depicts the main phases in a criminal investigation. There are four phases that make up this framework: the readiness phases, the deployment phases, the presentation phases and the digital investigation phases. The readiness phases involve the training of required people and the testing of tools and utilities that will be used in the investigation of the system. This set of processes is known as the operations readiness phase. The infrastructure readiness phases revolve around the configuration of appropriate equipment (Carrier and Spafford, 2004).

The deployment phases are made up of two sub phases; these are the detection and notification phase and the confirmation and authorisation phase. The detection and notification phase takes place when the incident occurs and the investigators are alerted. The confirmation and authorisation phase involves the investigators being authorised to start the investigation process.

Figure 2.22 shows that a central piece of the framework is that of the physical crime scene investigation phase. This phase can only begin once the investigators have been authorised to begin the investigation. Here physical aspects are taken into account and physical evidence is gathered. Another important factor in this step of the investigation is that certain physical aspects of the crime scene can be reconstructed. Once the digital physical objects have been seized the next step begins, known as the digital crime scene investigation phases. In this step the devices are examined for digital evidence; Carrier and Spafford, (2004) state that this phase comprises the

preservation of the system, search for evidence on the acquired digital devices and the reconstruction of the digital events.

The presentation phase is the final phase in this framework. It is here that the theories and evidence will be presented to the appropriate authority in a manner that is legally acceptable.

### 2.13 Digital Crime Scene Investigation Framework

It is important to note that the crime scene investigation framework discussed above (figure 2.22) deals with any crime and not just a digital one in particular. This being said, Carrier and Spafford (2004) have developed a framework specifically for the investigation of digital crimes.

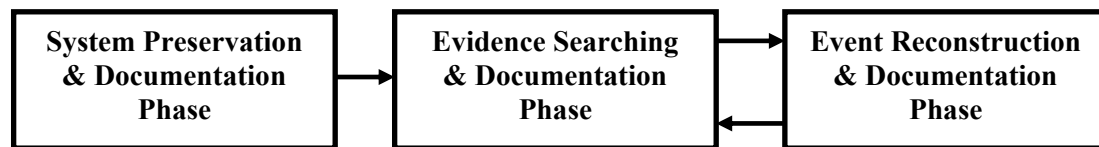


Figure 2.23: The Digital Crime Scene Investigation Phases, from: Carrier and Spafford (2004)

Figure 2.23 depicts the digital or cyber forensics investigation framework. According to Carrier and Spafford (2004), this model consists of three phases: the system preservation and documentation phase, the evidence searching and documentation phase and the event reconstruction and documentation phase. The system preservation and documentation phase involves the imaging of any digital evidence as well as the documenting of all relevant case material. The evidence searching and documentation phase involves the extraction of data from these images made in the previous step. The event reconstruction and documentation phase, is simply stated as the reconstruction and simulation of the crime scene in order to gain a better understanding of the events that have occurred. These three phases are further discussed in the following sections.

#### 2.13.1 The System Preservation and Documentation Phase

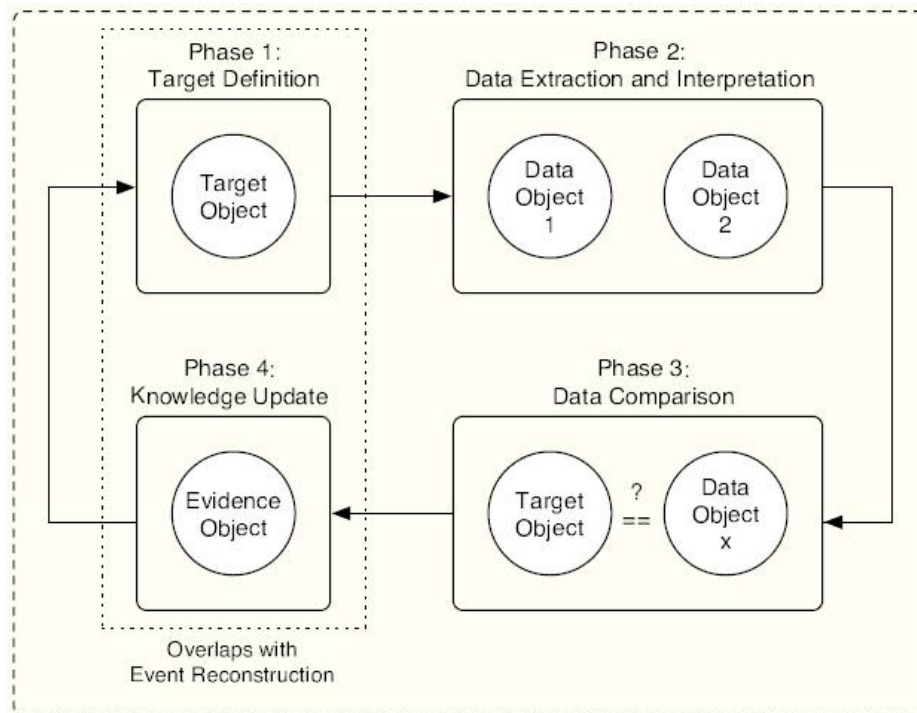
This phase deals with the “taking down” of digital devices and machines in such a way as to preserve the chain of evidence and to ensure that none of the original evidence becomes changed or altered in any way. Carrier and Spafford (2004) notes that some events that could take place during the course of this phase include both the

imaging of systems, as well as the duplication of hardware. In essence, this phase involves the preservation of the crime scene while seizing digital evidence. Bunting and Wei (2006) reiterate the importance of preserving the crime scene and making a copy of the original data known as an image. They add that for this image to be the legal duplicate of the original data, every one and zero (binary) must be an exact duplicate of the original. All electronic data are made up of ones and zeros; this is the most basic form in which electronic data can be represented. Bunting and Wei (2006) describe the following steps in the forensic investigation first response: search authority, securing the scene, recording and photographing, shutting down the computer, and bagging and tagging. This first step of search authority refers to the necessary permission to perform the search and seizure. Securing the scene is referred to as the most important aspect as it pertains to the safety of the investigators on the scene. It is noted that a security team should first establish a safe perimeter and make initial entry. Recording and photographing the scene is vital and should be performed before anything is touched. Shutting down computer systems must adhere to the correct procedures for bagging and tagging computers that are online or powered on. One cannot simply just unplug computers as valuable data may be compromised or even destroyed. For example, before pulling the computer's power plug out of the wall socket (as is required for taking down a machine), depending on the operating system, investigators may need to take a photograph of the screen. Bunting and Wei (2006) state the way in which a computer should be taken down depends on the operating system it is running, and for many Windows based computers, it is adequate to simply pull the plug out of the wall socket. Bagging and tagging involves the cataloguing and storing of digital evidence and can only be done once the computers are shut down. Bagging prevents tampering of the evidence and tagging is the adding of valuable crime scene information to the evidence e.g. time, date and location. Tagging also helps preserve the chain of custody. Bunting and Wei (2006) explain the chain of custody concept to be that of the handling of evidence to maintain its integrity and remain admissible in a court of law.

### **2.13.2 The Evidence Searching and Documentation Phase**

This phase takes place once the crime scene is secured and the evidence has been properly tagged and bagged. It involves the extraction of data from the evidence

and is a four phase process, as depicted in the diagram below (Carrier and Spafford, 2004).



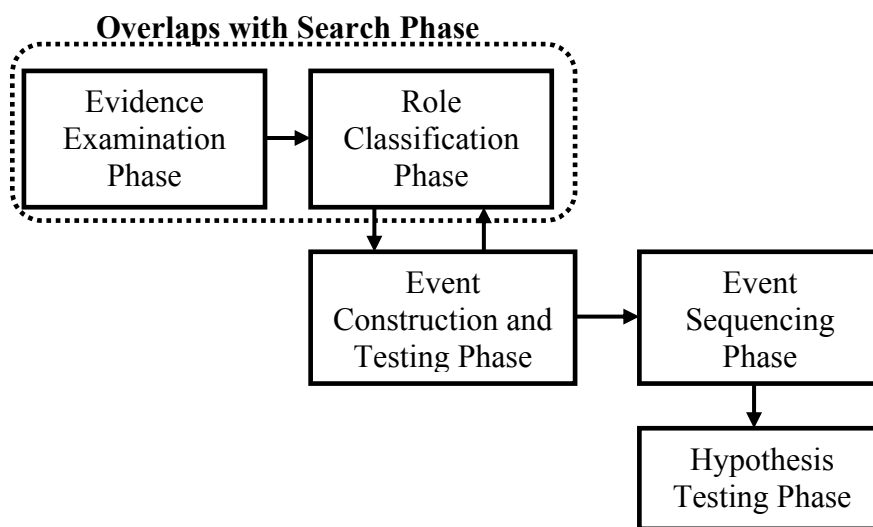
**Figure 2.24: The Evidence Searching Phases, from: Carrier and Spafford (2004)**

Figure 2.24 shows the four phases with regard to evidence searching. Defining targets for the investigation process is a difficult task that must be done before the other phases can begin. These targets can be made up from the investigators' past experience and even characteristics of digital evidence e.g. hidden files. Once the targets are defined and a list of "keywords" has been created, the data extraction process can commence. This keyword list is used to search for data on a suspect's hard drive or digital data store; this is often an automated process using forensic software. The investigator will also search through the results as certain keywords may bring up excessive amounts of data; for example, if one is searching for the word "loan", it could bring up hundreds or even thousands of results. Should this be the case, reduction techniques must be used to sift out the pertinent data (Carrier and Spafford, 2004). Here the data comparison phase comes in as the results from the searches have to be compared to what is required to ensure prosecution. The final phase, known as knowledge update, pertains to the correct storing of the acquired evidence. Here the MD5 and CRC values (as discussed in 2.11.2) are used to preserve the integrity of the data in order to ensure admissibility in a court of law.



### 2.13.3 Digital Event Reconstruction

This is the final phase in the digital forensic framework and deals with the circumstances surrounding how the elements found during the searching phase came to be. According to Carrier and Spafford (2004), this phase attempts to answer the question of what events the acquired evidence was involved in and what events occurred due to this object taking place. Carrier and Spafford (2004) describe five sub-phases that occur during the digital event reconstruction. These are the evidence examination phase, role classification, event construction and testing, event sequencing phase and lastly, the hypothesis testing phase.



**Figure 2.25: Evidence Reconstruction Phase, from: Carrier and Spafford (2004)**

During the evidence examination phase, the individual pieces of evidence (objects) are identified and categorized according to their characteristics. The role classification phase creates theories regarding what function and role the objects could have played. These two phases are partially touched on during the previous search phase. The event construction and testing phase uses the cause and effect relationship of the object's roles created in the previous phase, to create events that could have occurred. The event sequencing phase is simply the organising of the events that were reconstructed into a sequence of occurrences. Lastly, hypothesis testing is the step where conclusions about the events are tested to ensure consistency.

For the purposes of this study, a conceptual framework will be created using elements of Carrier and Spafford's (2004) digital investigation framework from a

global perspective. The new framework will be a conceptual one that will be used specifically for the digital forensic auditing of documents.

### 2.14 Cyber Tools online search for Evidence (CTOSE)

CTOSE developed a reference process model as a guideline as to how a company should proceed when a computer incident occurs. The model focuses on how digital data should be acquired, collected, stored, secured and analysed in order to be legally admissible in a court of law (Broucek and Turner, 2004).

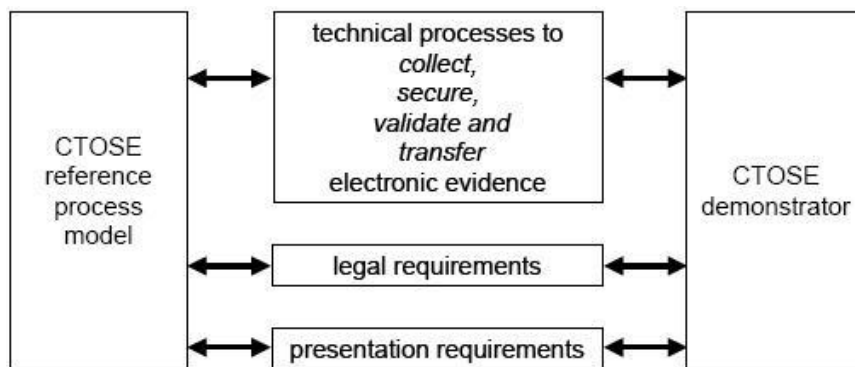


Figure 2.26: CTOSE Reference Process Model, from: Broucek and Turner (2004)

The reference process model contains five phases. These are the preparation, running, assessment, investigation and learning phases. The model depicts the actions and decisions that have to be conducted as part of an investigation into computer misuse.

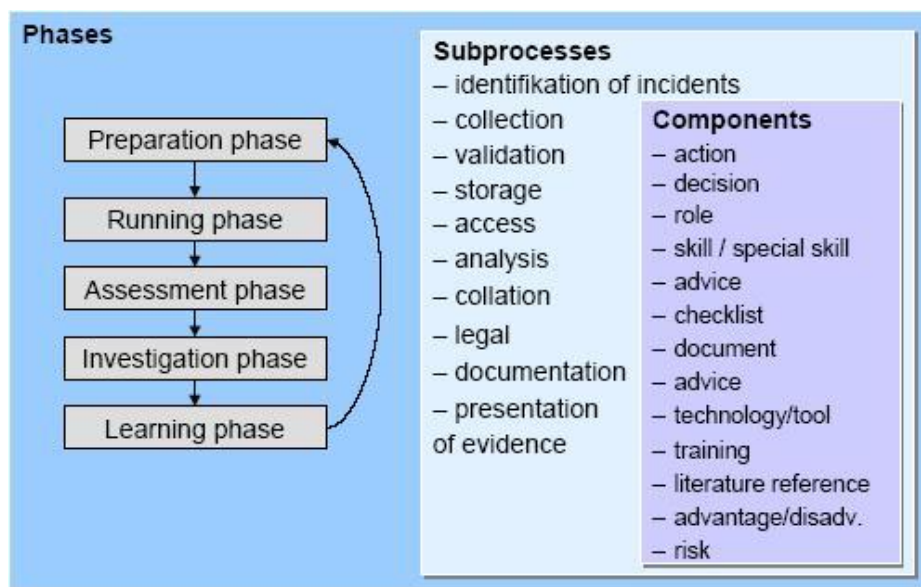


Figure 2.27: CTOSE Phases of Response, from: Broucek and Turner (2004) (Contents SIC)

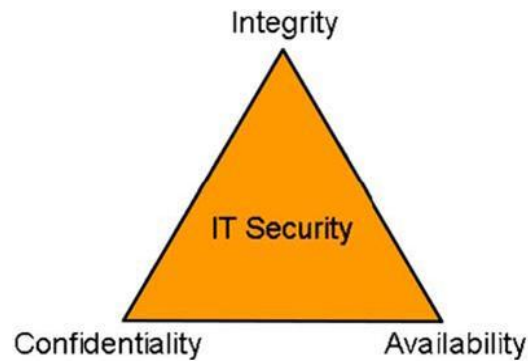
Figure 2.27 shows that at the end of the five phases, the process reverts to the preparation phase as critical (the most important) information technology security measures are defined and implemented here. The diagram also depicts sub-processes and components, which run hand-in-hand with each other.

## **2.15 Digital Forensic Framework that incorporates Legal Issues (FORZA)**

FORZA is the shortened term given to the Forensics Zachman Framework. According to Jeong (2006), many digital forensics processes and tasks are based on procedures created by traditional forensics scientists. It is said that these scientists were very focused on the procedures in handling and capturing the evidence. This has resulted in forensics practitioners of today following in their footsteps and forgetting that the legal aspects play an equally pivotal role as any other process in the digital forensics investigation. Jeong (2006) adds that technical barriers exist between information technologists, legal practitioners and investigators; therefore, in order to get past this problem, a technically independent framework needs to be developed. It has often been the case where an electronic forensic case would go to court and would lose simply due to the fact that legal technicalities during the process of the investigation were not adhered to e.g. the contamination of a crime scene simply by having a person not directly involved in the investigation on the accused's premises. Jeong (2006) states that a wide gap exists between technical specialists and legal practitioners. Furthermore, he states that they do not need a certain level of understanding to delve deeply into a computer's architecture; they simply need to know whether the data is necessary to the case and will not get thrown out of court.

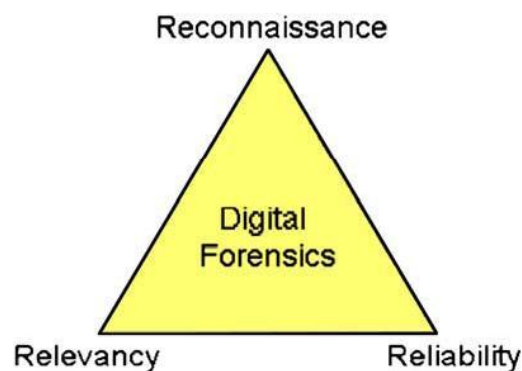
### **2.15.1 Fundamental Principles in Digital Forensics Investigation Procedures**

Many different technological aspects exist in the IT (Information Technology) security field, each with its own set of principles. These technologies include items such as biometrics, card readers, scanners and specialised security systems. However, all these different technological aspects rely on a single set of fundamental principles, namely: integrity, confidentiality and availability. These three core aspects link the different areas of IT security together (Jeong, 2006).



**Figure 2.28: IT Security Fundamentals, from: Jeong (2006)**

This being said, a digital forensics investigation should have its own set of core principles. According to Jeong (2006), there exist three fundamental principles in the digital forensic investigation. These are reconnaissance, relevancy and reliability. These three principles are pivotal in order to relate extracted information from a digital crime for admissibility in a court of law.



**Figure 2.29: Digital Forensics Investigation Fundamentals, from: Jeong (2006)**

Reconnaissance refers to the actions that need to be taken before ethical hacking of the evidence can take place. This essentially means that all methods, procedures and tools need to be utilised on the data, no matter where it is stored, in order to discover, extract and analyse for instance. It is important for digital forensic investigators to focus on the truth behind the data (Jeong, 2006). The reliability factor refers to the extraction of the data in a way that would not compromise its admissibility in a court of law. Relevancy refers to the following question: if the evidence were to be admissible in a court of law, would it be relevant to the case? This factor delves mainly into the usefulness of the evidence, as it is unethical to take information that is irrelevant to the investigation.

### **2.15.2 The Participants in the FORZA Framework**

There are many different participants in a digital forensic investigation, each with their own set of responsibilities and duties. A typical investigation would involve the following role players: the case leader, system/business owner, legal adviser, security/auditor, digital forensic expert, digital forensic investigator/operator, digital forensic analyst and the legal prosecutor (Jeong, 2006). The case leader is the person who plans and leads the case as well as the steps in the forensic investigation. The system/business owner is the owner of the investigated system. This person is usually the victim or sponsor of the investigation. The legal adviser is the participant on whom the case leader relies for legal advice, as this participant advises whether it is feasible to move forward and what procedures are necessary for the collection/seizure of evidence. The legal adviser can also save large sums of money by identifying whether or not it may be feasible to proceed with the investigation at an early stage in the investigation process. The system security/auditor exists in medium to large organisations. These people should be interviewed by the case leader so as to ascertain the scope of the case and exactly what methods of design were used in the system. The digital forensic specialists can be hired by the case leader to plan the investigation strategy and to decide if it is feasible to hire expert consultants for specific parts of the investigation. The results generated by these digital forensic specialists would then be given to the digital forensic investigator. This participant's function is to collect, extract, preserve and store the digital data from the systems. Once this evidence is collected, the digital forensic analyst would extract the relevant data for the investigation. Various tests or procedures may have to be run on the data as well as the reconstruction of the timeline of the case. Once all the relevant information has been extracted, the case leader would approach the legal practitioner to determine the feasibility of proceeding with the case.

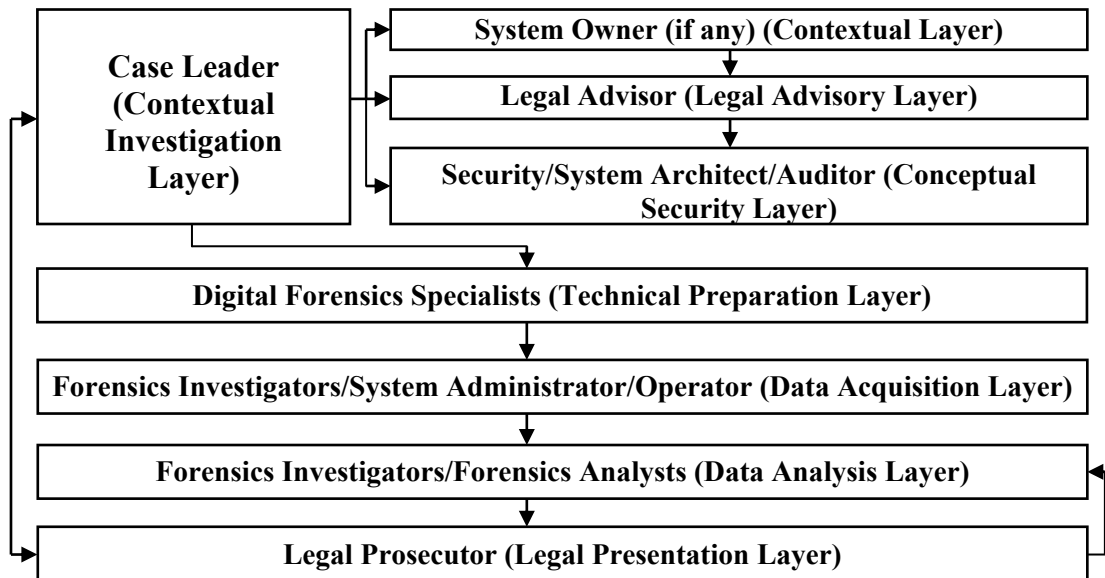


Figure 2.30 Process flow between the roles in digital forensics investigation, from: Jeong (2006)

This diagram clearly illustrates the various role players and their functions within the investigation process. It is important to note the sequence in which the steps flow. Jeong (2006) points out that all the layers of the investigation model are connected to one another through six questions.

### 2.15.3 The Six Key Questions

According to Jeong (2006), the FORZA framework is essentially the Zachmans framework incorporated within six key questions. These questions are: the what (data attributes), why (motivation), how (the procedures), who (the people), where (the location), and when (the time). These questions link the various layers and participants of the framework together. A diagram of these questions applied to all the various participants in the framework is shown in Table 2.4 below:

	Why (motivation)	What (data)	How (function)	Where (network)	Who (people)	When (time)
Case leader (contextual investigation layer)	Investigation objectives	Event nature	Requested initial investigation	Investigation geography	Initial participants	Investigation timeline
System owner (if any) (contextual layer)	Business objectives	Business and event nature	Business and system process model	Business geography	Organization and participants relationship	Business and incident timeline
Legal advisor (legal advisory layer)	Legal objectives	Legal background and preliminary issues	Legal procedures for further investigation	Legal geography	Legal entities and participants	Legal timeframe
Security/system architect/ auditor (conceptual security layer)	System/Security control objectives	System information and security control model	Security mechanisms	Security domain and network infrastructure	Users and security entity model	Security timing and sequencing
Digital forensics specialists (technical preparation layer)	Forensics investigation strategy objectives	Forensics data model	Forensics strategy design	Forensics data geography	Forensics entity model	Hypothetical forensics event timeline
Forensics investigators/system administrator/operator (data acquisition layer)	Forensics acquisition objectives	On-site forensics data observation	Forensics acquisition/seizure procedures	Site network forensics data acquisition	Participants interviewing and hearing	Forensics acquisition timeline
Forensics investigators/ forensics analysts (data analysis layer)	Forensics examination objectives	Event data reconstruction	Forensics analysis procedures	Network address extraction and analysis	Entity and evidence relationship analysis	Event timeline reconstruction
Legal prosecutor (legal presentation layer)	Legal presentation objectives	Legal presentation attributes	Legal presentation procedures	Legal jurisdiction location	Entities in litigation procedures	Timeline of the entire event for presentation

**Table 2.4: The six questions applied to all participants in FORZA, from: Jeong (2006)**

#### **2.15.4 Implementing the Legal Aspects**

The FORZA framework allows for different standards and procedures to be linked together. This means that the digital forensics investigation will no longer be a fully technical process, as by incorporating the legal aspects into the processes of the investigation, the chances of losing the case due to investigative mistakes is substantially reduced. Additionally, this framework gives the legal adviser a chance to determine early on in the investigation if it is feasible to continue the investigation process, thereby potentially saving the client large sums of money.

This being said, the legal adviser and prosecutor can focus on the following questions adapted from Jeong (2006):

##### **The legal adviser's questions:**

- Legal objectives (Why)

The legal adviser will ask what the purpose of the dispute is, what the law of the dispute is and whether the case is a criminal or civil case.

- Legal background and initial issues (What)

Here the adviser will ask what the relevant laws are, which sections of the relevant laws should be referred to, what data should be collected and lastly, what information is relevant and required.

- Legal procedures (How)

This refers to aspects such as whether any search warrant is necessary or if any steps need to be taken for the protection of the collected evidence.

- Legal geography (Where)

This question is extremely important if crimes are committed across countries borders. If this is the case, the necessary laws will have to be researched and taken into account as legal rules are applied very differently in various countries.



- Legal entities and participants (Who)

The questions of who the respondents and claimants are as well as who the legal staff are, must be raised.

- Legal timeframe (When)

The time limits of the case need to be determined here as well as any other time related details regarding the duration of the case. Research also needs to be done in order to determine what the usual time and costs of similar cases amounted to.

Once all these questions are answered by the legal adviser, the case leader can utilise this information and retrieve the relevant evidence. Subsequently, once the forensic analysis is complete, the case leader will be able to construct a review with the legal prosecutor.

#### **The legal prosecutor's questions:**

- Legal presentation objective (Why)

An important question to start off with is whether or not it is feasible to proceed with the case. The prosecutor will also have to determine if sufficient evidence has been acquired and which litigation mechanism should be used.

- Legal presentation attributes (What)

Here the question of what charge should be issued needs to be determined. The evidence that pertains to the case needs to be ascertained as well as what evidence is going to be presented in the court of law.

- Legal presentation procedures (How)

What tactics/schemes should be used during the trial, is of relevance here.

- Legal jurisdiction location (Where)

The questions of where the hearing is going to be held needs to be answered, as well as any other questions regarding the location of enforcement and litigation.

- Entities in the litigation procedure (Who)

Which witness/witnesses should be called upon during the trial, must be addressed. Questions as to whether or not it would be feasible to use expert witnesses during the trial need to be determined, and information as to which judge, councillor and arbitrators are involved, needs to be acquired.

- Timeline of the entire event for presentation (When)

Here the entire crime events need to be reconstructed. Checks need to be done to make sure there are no timelines missing in the evidence. The question of when the case should be presented needs to be determined.

Once all these questions have been applied to all participants in the framework, the case leader can be confident that the legal aspects have been adhered to and correctly analysed. These questions will work hand-in-hand with most forensic frameworks during the running of an investigation to determine authorship of a document.

## **2.16 A Conceptual Framework for the Forensic Auditing of Academic Assignments**

The process of designing a conceptual framework is a time-consuming one. Frameworks are especially important in assisting the researcher to more effectively solve the task at hand: this is complemented by the fact that frameworks contain and utilise the ideas and knowledge of others. The systematic method of approaching problems using a framework enables the researcher to decide on a variety of different methods to use when solving a problem. This study shall involve the analysis of various different frameworks, then grouping the pertinent factors together and creating a new conceptual framework specifically designed to address the problem

that this research aims to solve. According to Mujer Sana (2003), the following points embody a conceptual framework:

- i. A set of ideas grouped in such a way as to make them easier to communicate to others
- ii. A structured process of logic which allows for greater understanding of the project's tasks
- iii. The research and work of other individuals that gives the researcher further understanding and knowledge regarding the project at hand
- iv. A clearer understanding of the subject matter with regards to the task at hand
- v. A set of assumptions, values and definitions to assist in the tasks at hand

### **2.17 Content Analysis**

Stemler (2001) cites Holsti (1969) in explaining that content analysis can be defined as “any technique for making inferences by objectively and systematically identifying specified characteristics of messages”. Stemler (2001) adds that this definition does not simply stop at text analysis, but extends to that of picture analysis and video tape analysis. In addition, Stemler (2001) notes that a good technique to use when determining authorship is a method known as function word analysis. This is made possible by examining a person's prior writings, allowing for a frequency of function words to be created and used to determine the probability of authorship for the text in question. Content analysis enables the investigator to sift through a large volume of data with ease and in a systematic fashion (Stemler, 2001 cites GAO, 1996), as well as being extremely useful in analysing trends or patterns in documents.

During the phase where documents are being collected for content analysis, three possible problems could arise. According to Stemler (2001), should a significant number of documents from the total population be missing, then the content analysis would be impossible to complete. The second possible issue that could arise is the archiving of inappropriate records; it is imperative to identify these and discard them. The final problem one could encounter is that of uncodable documents. Uncodable in this regard is taken to mean those documents that have irrelevant content or missing pages and sections, for instance. According to Stemler

(2001), three types of units are available for content analysis. These are the sampling, context and recording units.

### **2.17.1 Sampling Units**

Sampling units embody a number of different things that are dependent on how the researcher derives meaning. Meaning in this context is simply what the researcher deems critical to the subject that he/she wishes to research.

### **2.17.2 Context Units**

These set the physical limits on the data that the researcher is trying to record. In this instance they could be sentences, paragraphs or even entire portions of a document.

### **2.17.3 Recording Units**

These units do not have any physical boundaries and represent the ideas behind the statements.

## **2.18 Character Recognition and Plagiarism**

Character recognition was at one stage a rather costly system for any company to use. However, with advances in technology, mainly that of scanners and computer software, many companies can now afford to implement these systems. Character recognition software allows the user to scan a hardcopy page of text and/or pictures and have it transformed into editable electronic format on one's personal computer. This has many advantages, but by the same token the system also enables abuse and plagiarism of written material. In order to understand this issue, one must distinguish between handwritten characters and that of computer typed characters. Handwritten characters are far more difficult to convert into electronic text as opposed to simply scanning typed text. Thus, in order to convert the handwritten text, one would also need very sophisticated optical character recognition (OCR) software, allowing us to conclude that plagiarising a handwritten book is far less likely than its typed counterpart. The OCR process is by no means a simple one, as stylistics and character recognition have to work differently in order to interpret the symbols. With stylistics one can compare an accused offender's handwriting with that of the offending material in order to establish if it was indeed the accused that had written the material in question. However, character recognition software works very differently in order to interpret handwritten notes with those typed on a computer.

The natural handwriting recognition (NHR) process allows computers to read and recognise handwriting with a high degree of accuracy (Rawso, [no date]). NHR is a series of tasks that is outlined below.

### **2.18.1 Form Identification**

This stage is the identification of the form image presented for character recognition. The output of this step is either that the form is unrecognisable or that it is acceptable for further processing.

### **2.18.2 Field Isolation**

Here the text images for each data field are extracted from the form. The output here is the text without the surrounding empty portions of the form.

### **2.18.3 Segmentation**

This process pertains to the breaking of each image of text from the field isolation step, into small images for processing.

### **2.18.4 Recombining Segments**

This entails identifying various combinations of segments as possible matches for isolated character images. The output here is in the form of isolated character images.

### **2.18.5 Recognition**

This phase assigns confidences to all allowed classes for each character image candidate.

### **2.18.6 Organising Character Candidates**

This step organises the output of the previous two steps into a usable form for dictionary input.

### **2.18.7 Dictionary Based Correction**

This process selects the dictionary entries that best match the output from the previous step (i.e. the organised character candidates). The output here is the best match for the scanned image of the text.

### **2.18.8 Levels of Acceptance**

This step compares the dictionary image selection to set accuracy levels. The output here will be an acceptance or rejection of the final result as a conversion of the written data.

### **2.18.9 Rejection of the Result**

Should the system detect a discrepancy in the output, it would send the image for human correction or acceptance.

A basic summary of the NHR system would be that it selects a word from the dictionary and assigns it a confidence rating. Should the confidence rating not be high enough, the rejected answer can be sent to the user for confirmation. Manual input can also be used for words that are only comprehensible by humans. According to Rawson (no date), at even a 50% recognition ratio, NHR cuts down the human work by 50%. A study conducted by Rawson [no date] used the National Institute of Standards Database which contained real words written by five hundred independent writers. The data was then inputted showing a 98% recognition rate.

According to Rawson [no date], one of the most advanced products in circulation today is ParaScript. NHR can be a most powerful tool in the war against plagiarism. The ability to simply feed handwritten data into a scanner and have it automatically converted into computer text is a great step, as it reduces the time required if one were to manually re-type every word into a computer. Once the handwritten data (e.g. from assignments) is entered onto the computer, they can then be submitted to any of the anti-plagiarism tools (e.g. *Turnitin.com*) in existence, in order to be cross-referenced for plagiarism.

### **2.18.10 Summary**

NHR techniques play an important role in determining authorship, as it facilitates for quick and efficient transcoding of written text into an electronic format. Once text is in electronic format, a number of corpus analysis techniques can be performed on the data. A variety of software tools have been created for corpus analysis, these tools all have one fundamental requirement; that the data being analysed is in electronic format. NHR techniques will make it possible to move from manually performing analysis on large hand written corpora to a quick and efficient electronic system of analysis. It is important to note that NHR is not an anti-

plagiarism tool itself, but rather one that can be used in conjunction with corpus analysis methods.

## **2.19 Anti-Plagiarism Tools**

A number of tools are in existence to combat plagiarism. These tools can be broken down into one of two categories – Online tools and off-the-shelf software packages. Please refer to sections 2.6.7.1 and 2.6.7.2 for specific examples regarding some of the available packages and their respective features.

### **2.19.1 Online Tools**

These particular tools are classified as such due to the fact that they only exist on the Internet and cannot be downloaded onto a user's personal computer. These systems are located on secure computer servers. In order to use the facility the user will have to open an account with the service provider e.g. *Turnitin.com*. Once the user has an account they can then upload quantities of computer text to the server; which will then attempt to find similarities against its own database of literature as well as Internet searches. Google is an example of one of these so called online tools. Although its function is primarily that of a web search engine, many individuals, organisations, among others use it for the detection of plagiarism (BMJ, 2006).

### **2.19.2 Offline Tools**

These tools can be purchased and installed onto a user's computer. From here it is a simple matter of feeding in computer text into the program that will attempt to find similarities in one or more documents of the user's choice. Some software packages also allow for finding similarities in Internet documents by using automated searches.

## **2.20 Conclusion**

This chapter delved into the theoretical basis for the aspects that are central to the research questions in this study. In order to understand the concept of cyber forensic auditing with regard to ethical writing, one needs to gain an understanding of the entire body of knowledge surrounding these concepts. To begin with, this study first aims to give an understanding of ethical writing, what it is and what it entails (section 2.4). It then goes on to give an understanding of cyber forensics and shows how the related terms (e.g. electronic forensics) are linked to it and have led to its

development. Tools to aid in the fight against plagiarism are also reviewed. Once the concepts are put into perspective, a suitable forensic framework needs to be analysed and adapted to this particular form of investigation. Sections 2.12 to 2.15 explored the usage of frameworks: the criminal investigation framework is used as a starting point, thereafter the digital investigation framework and the FORZA framework are described. The research also takes a deeper look into more complex electronic forensic methods, terminology and processes (e.g. MD5 and CRC).

Section 2.3.3 presented an introduction to word frequency analysis in order to give a basic perspective on the relevance of this system in the fight against plagiarism. Software packages that could be used to assist in this endeavour can also be of great value. These include both online (e.g. *Turnitin.com*) and downloadable programs (e.g. SPLAT). Section 2.17 described a system known as function word analysis and how it fits into the greater system of content analysis that forms part of one of the key underlying concepts of this study, known as corpus analysis.

The next chapter will examine the analysis of the collected research data as well as provide the theoretical basis and reasoning for the chosen methods of analysis.



## *Chapter Three*

### FORENSIC LINGUISTICS

*It is important to note that speakers and writers do what comes naturally: they quickly acquire the ability to construct grammatical and acceptable syntactic structures, then they produce utterances which are more or less elaborate sentences. Linguists find ways to observe and understand how language is acquired, analyze what speakers and writers do, and then account for how and why.*

McMenamin (2002)

#### **3.1 Introduction**

This chapter will closely follow McMenamin (2002) as a major source, chiefly because information on the issue of forensic linguistic data is very scarce. This study approaches grammar from the theoretical perspective of Quirk et al. (1974). In this thesis, forensic linguistics will only be discussed with regard to syntax (sentence patterns) and the lexicon (word categories, the words that belong to them and the frequency at which the words are used in spoken text). The thesis will not deal in detail with aspects of language use such as pronunciation (phonetics), sound patterns (phonology), word formation patterns (morphology), stylistic differences in language use, and sociolinguistics (what language use reveals about the social status of users). Phonetics or phonology, stylistics and sociolinguistics will be summarized from McMenamin (2002), who presents an excellent review of these aspects of forensic linguistics.

Finally, the term “idiolect” is considered important, because it relates to aspects of language use, characteristic of a particular individual. This enables the forensic linguist to establish the probability of authorship in texts where authorship is in question.

According to McMenamin (2002), linguistics is the scientific study of language and language is the combination of words and sounds among other factors, which is generally understood by its community. He also notes that forensic linguistics is but a part of the greater field of applied linguistics. The patterns, according to which words,

sounds and sentences are formed, are collectively known as *grammar* in linguistics, the field of scientific language study. This thesis takes the grammar of Quirk et al. (1974), as a point of departure. Section 2.3 in the previous chapter, on “The Role of Language Analysis in Forensic Auditing” presented general outlines of the aspects of linguistics that pertain to forensic linguistics, and forensic linguistics as part of forensic auditing.

According to McMenamín (2002), language is a system of communication that humans use to convey messages. Sounds are combined (words/utterances) to create meaning and sentences which results in what is known as natural language. Language usage can fall into two categories, “form” and “function”. McMenamín goes on to add that language “form” relates to the linguistic structure of language usage, while “function” refers to the actual usage of language as part of social interaction. Linguistic experts have broken language usage down into several categories, which allow for more effective analysis and a better understanding of the language. McMenamín (2002) describes six types of language forms. These are as follows:

- Phonetics – the study of speech sounds as well as the way in which these sounds are created by the human body
- Phonology – the study of how sounds in a particular language follow predictable patterns, e.g. cat, bat and hat
- Morphology – the study of how words are created and formed using various sounds
- The Lexicon – the study of how words and word parts combine into larger units, for example sentences and clauses
- Syntax – the study of how words are combined into longer sentences and sequences
- Semantics – The study of the intended meaning derived from word usage

For the purposes of this study, the research will focus mainly on lexicon and syntax, as it is these that play the most prominent part in the forensic stylistic investigation. The lexicon is the area of the language that holds all the different words, therefore the lexicon of a particular language is its dictionary and the lexicon of a particular person is that person’s word knowledge (McMenamin, 2002). Syntax

is focused mainly on how words are combined to form longer sequences i.e. sentences and phrases among others. He adds that most sentences follow a two phase description. The first is the structure of the phrases and the second refers to how language users manipulate these phrase structures in order to create new and even more complex sentences. McMenamin (2002) makes an important point, that linguistics is a “descriptive” process rather than “prescriptive”. This means that linguistics embodies a process of understanding language and its usage, rather than it being “prescriptive” where the goal is to create rules for appropriate language usage.

SENTENCE						
NOUN PHRASE		AUXILIARY VERB			VERB PHRASE	
(ART) (ADJ) NOUN SENT	MODAL	PERF	PROGRESSIVE	MAIN VERB	COMPLEMENT	ADVERBIALS
John				snores.		
Mary				slept		there.
The man	might			stop		here.
A tall student			is	studying	math	in the classroom.
His sons	may	have		driven	the car	carefully.
A good teacher	should		be	thinking		clearly every day.
These two noisy birds	could	have	been	looking for	other birds	all night long.
Many students	will			seem	quiet	at first.
Mom's car			was	rolling		down the street.
Her sister				rolled down	the shade	quietly last night.
The player who got hit	must	have	been	hoping	to draw a foul	during the game.

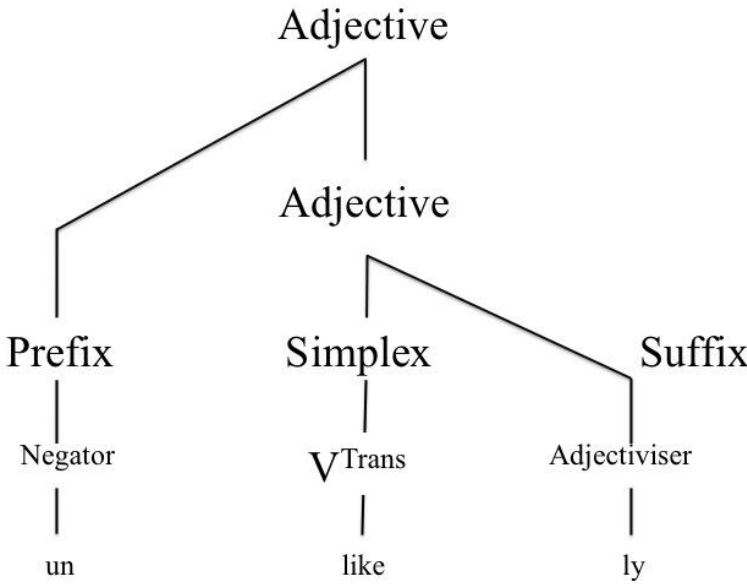
**Table 3.1: “Basic phrase structures of English”, from: McMenamin (2002)**

The table above clearly depicts the typical sentence structure with its lexical categories, similar to that of the transitive sentence patterns that were analysed in section 2.1 (page 22).

TRANSFORMATION	TYPE	BASIC PHRASE STRUCTURE	CHANGE TO NEW STRUCTURE
<b>Negation</b>	Addition	John may go.	John may <u>not</u> go.
<b>Yes/No Question</b>	Movement	John will go.	<u>Will</u> John go?
<b>Contraction</b>	Deletion	John should not go.	John <u>shouldn't</u> go.
<b>Do insertion</b>	Addition	John loves Mary.	John <u>does</u> love Mary.
<b>Negation</b>	2 additions	John loves Mary.	John <u>does not</u> love Mary.
<b>Wh- Questions</b>	All 3	John will leave now.	<u>When will</u> John leave [ <u>   </u> ].
<b>Passive</b>	All 3	John saw Mary.	<u>Mary was</u> seen.

**Table 3.2: “Commonly used transformations of English”, from: McMenamin (2002)**

Transformations are simply the effects that cause a change in sentence structure. The table above depicts some of the various transformations that are in existence. This analysis will assist in a linguistic forensic audit as the usage of these transformation categories could be attributable to a person’s writing style.



**Figure 3.1: The negative transformation in the word “Unlikely”**

The diagram above clearly indicates the negative transformation in the word “unlikely”. This type of transformation falls under the category of “Negation”.

McMenamin (2002) notes that forensic linguistics is not a new form of forensic investigation. He goes on to add that language users have both productive and receptive capabilities. The productive capabilities are comprised of speaking and writing and the receptive capabilities are made up of listening and reading.

Linguistic Level	Spoken Language	Written Language
<b>LANGUAGE FORM</b>		
PHONETICS	Sounds ( <i>phonemes</i> )	Letters ( <i>graphemes</i> )
PHONOLOGY	Sound patterns + blends	Letters + digraphs
MORPHOLOGY	Word formation ( <i>morphemes</i> )	Word parts: ( <i>roots + affixes</i> )
LEXICON	Words	Vocabulary ( <i>dictionary</i> )
SYNTAX	Sentence formation	Written sentences ( <i>grammar</i> )
SEMANTICS	Expressed meaning	Meaning of words + sentences
<b>LANGUAGE FUNCTION</b>		
DISCOURSE	Conversations + narratives	Written equivalents ( <i>stories</i> )
PRAGMATICS	Doing things with words	Written equivalents

**Table 3.3** “Linguistic levels in spoken and written language”, from: **McMenamin (2002, Section 2.1.2)**

Table 3.3 depicts the linguistic levels from least difficult and complex to most difficult and complex. The diagram shows the linguistic aspect followed by the two productive aspects known as spoken and written language.

### **3.2 Linguistic Analysis**

McMenamin (2002) states that linguistic theory, data and facts can be used to analyse language usage in greater detail. He also asserts that there are two forms of linguistics in existence, known as general and applied linguistics. General linguistics is more theoretical whereas applied linguistics is seen as being very practical, where its users place more importance on the reasons for observing the particular language. They are more concerned with the gathering of data and the analysis of their findings. He adds that linguistic analysis is a scientific process. McMenamin (2002) notes that there are many categories of linguistics in existence, for example, discourse analysis, language and gender, sociolinguistics and child language, among others. Forensic linguistics is but one of these mentioned categories; however, all the aspects of linguistics are used in forensic analysis. McMenamin (2002) states further that all areas of linguistics play a role in forensic application. It is important to note that not all of the following sections are central to this study: however, according to McMenamin (2002), they do form part of linguistic analysis and therefore are touched on briefly.

### 3.2.1 Auditory Phonetics

Auditory phonetics describes the study of language sounds and the interpretation of these sounds heard by the listener. It encompasses any methods of analysis that make use of auditory techniques in order to identify speakers by witnesses. According to McMEnamin (2002), forensic phonetics use auditory as well as acoustic methods of analysis. He adds that the primary concern of auditory analysis in forensic investigations is speaker discrimination, identification by witnesses and victims, and identification of class characteristics of speakers.

### 3.2.2 Acoustic Phonetics

This type of phonetics is the study of the physical characteristics of speech sounds as they leave the source and enter the air then eventually dissipate. Its method describes methods of speech analysis, using devices to identify suspects. According to Wapedia (2009), the field of acoustic phonetics was greatly enhanced by the invention of a device known as the “Edison phonograph” in the late 19<sup>th</sup> century. This device made it possible to record a conversation and later analyse it by replaying the recording and running it through different filters. This would yield a spectrogram of the speech utterances. This spectrogram is comprised of waveforms, which depict the amplitude (loudness), frequency (repetitions) and complexity of the sound being analysed. According to McMEnamin (2002), the primary concern of acoustic analysis for forensic investigations is that of speaker identification.

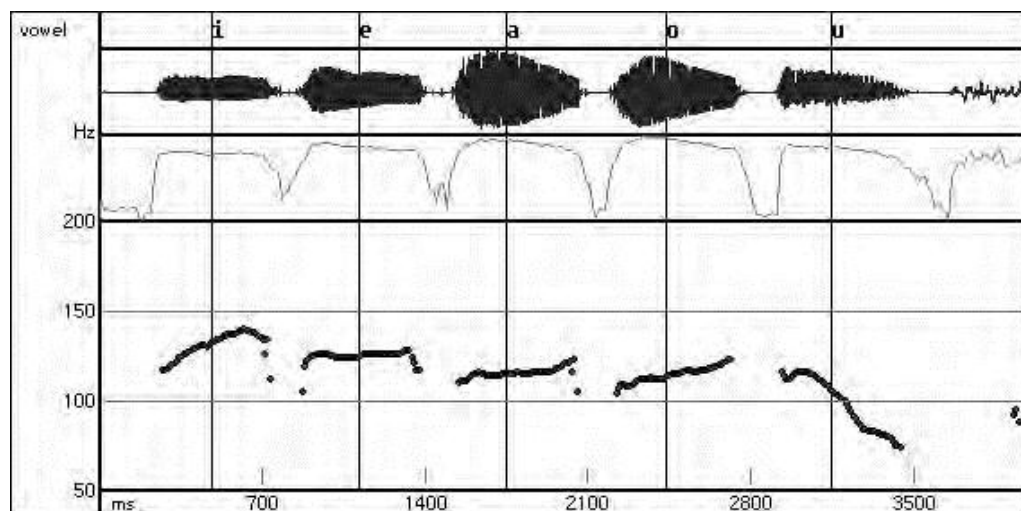
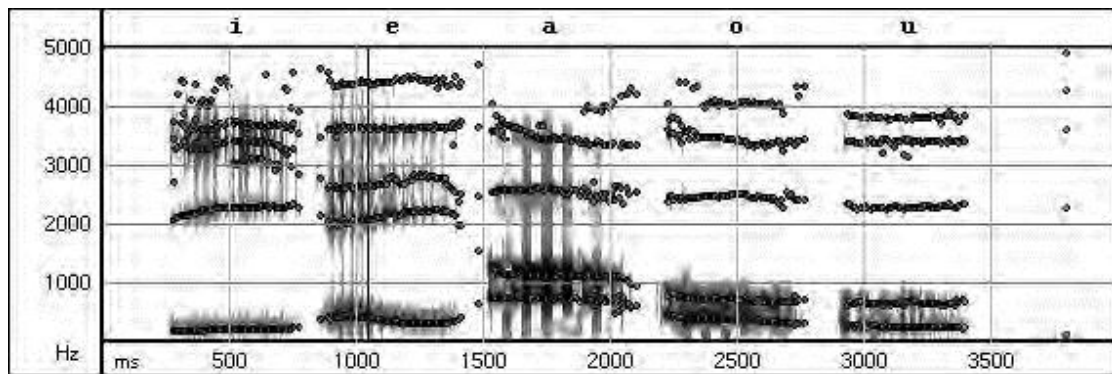


Figure 3.2: “Waveform, intensity contour, and fundamental for English vowels”, from: McMEnamin (2002, Section 4.2.2)

Figure 3.2 depicts a graph with waveforms for the English vowels *a, e, i, o, u*. The bottom X-axis describes the time in milliseconds taken to pronounce each letter. The Y-axis describes the frequency in Hz.



**Figure 3.3:** “Spectrogram for English vowels”, from: McMenamin (2002, Section 4.2.2)

Figure 3.3 depicts a spectrogram graph for the English vowels *a, e, i, o, u*. It is similar to the waveform graph above in that the milliseconds are on the X-axis and the Hz lies on the Y-axis. Each letter utterance is represented by a set of dots describing its waveform.

### 3.2.3 Semantics

This deals with the complexity of understanding written and spoken language that is difficult to understand, for example, disclaimers and legal acts. Certain documents could be written with jargon and discipline-specific terms; general language usage as well could be far more complex than that found in the average book. This is important as it will allow the investigator to determine if a particular document could be taken as being reasonably understandable by the average person.

### 3.2.4 Discourse Analysis

According to McMenamin (2002), this describes the study of language systems larger than sentences, for example, narratives and conversations. For investigative purposes, this deals with the analysis of conversations that occur within the judicial process and can be determined by variable factors such as a person’s social roles, topic, purpose and time.

### 3.2.5 Pragmatics

Pragmatics deals with the analysis of a speaker’s intended meaning, and uses both discourse analysis as well as pragmatics to analyse what context the language was used in, for example, conversations, dictation, promises and warnings.

McMenamin (2002) notes that pragmatics is important for forensic investigators as language users do not always convey their intended message with words on face value.

### 3.2.6 Stylistics

Stylistics pertaining to linguistic forensics plays three important roles. Firstly, it is used to determine if all the suspected writing were in fact authored by the suspect and secondly, if the writings in question were written by one of a number of possible authors. Lastly, it is used to determine if the material in question was authored by the suspected person based on external evidence. According to McMenamin (2002), linguistic stylistics can utilise two methods of authorship identification, namely, qualitative and quantitative methods. The qualitative approach is used when aspects of a language user’s writing is seen as being similar to that of another author. The quantitative approach is used when indicators are identified and then measured against occurrences in another text, such as that in table 3.4 below. McMenamin adds that these two methods are complementary and are often used together when determining authorship.

QUESTIONED Letter	KNOWN Suspect Writer #1	KNOWN Suspect Writer #2
Mary Ann	Mary Ann	Maryanne
Mary Ann	Mary Ann	Maryanne
	Mary Ann	Maryanne
	Mary Ann	Maryanne
	Mary Ann	Maryann
		Maryann

Table 3.4: “Spelling of “Maryanne” in QUESTIONED and KNOWN writings”, from: McMenamin (2002, Section 4.2.2)



### **3.2.7 Language of the Law**

This area of forensic linguistics is concerned with ensuring that legal documents, statutes, acts and other such documents are written in a manner that is clear and simple to understand (McMenamin, 2002).

### **3.2.8 Language of the Courtroom**

This is the analysis of the use of legal language by the various witnesses, judges and other legally inclined individuals. McMenamin (2002) notes that this area of forensic linguistics makes use of pragmatics and discourse analysis to affect the case outcome.

### **3.2.9 Interpretation**

This is the analysis of the interpretation from one language to another to determine its legal accuracy. This is used in situations where there is a question and answer system, which can occur during a trial proceeding. McMenamin (2002) states that interpretation is a very difficult task in forensic investigations as the translation of legal aspects is not as simple as just taking words on face value. The intended meaning is of paramount importance.

ELEMENTS OF SCIENCE USED IN LINGUISTICS	EXAMPLES FROM DIALECT VARIATION IN ENGLISH
1. Linguistic <i>facts</i> are observations of human language behavior.	A variety of English spoken by many African Americans demonstrates the forms in (1) to (4):  (1) I <b>am</b> goin'. (2) I'm goin'. (3) I goin'. (4) I <b>be</b> goin'.
2. <i>Classification</i> of facts results from ordering and grouping them.	First, sentences (1) to (4) fall into two groups, based on meaning. Second, sentences (1) to (3) are ordered "all to nothing": full verb, partial contracted verb, no verb.  <u>"I am going now."</u> <u>"I usually go."</u> (1) I <b>am</b> goin'.      (4) I <b>be</b> goin'. (2) I'm goin'. (3) I goin'.
3. <i>Relationships</i> are found based on the interaction of observed phenomena.	These facts interact with another set of facts (* sentences are ungrammatical):  (5) I know where I <b>am</b> . (6) * I know where I'm. (7) * I know where I.
4. <i>Principles</i> result from observing relationships among facts.	Two principles result from the facts observed in this variety of English. Principle #1: the verb <i>am</i> ( <i>be</i> ) can be contracted or lost before <i>goin'</i> . Principle #2: the verb <i>am</i> ( <i>be</i> ) cannot be contracted or lost at sentence end.
5. <i>Generalizations</i> result from combining various principles.	Principles #1 and #2 allow the generalization that deletion of the verb <i>am</i> ( <i>be</i> ) is related to contraction in this way: in this variety of English, <i>am</i> ( <i>be</i> ) is deleted only where it can be contracted in standard English.
6. <i>Theory</i> is based on the interrelatedness of generalizations.	This generalization and others like it allow hypotheses such as, African American English is rule governed, and it is related to other varieties of English in regular ways.
7. <i>Communication</i> and peer review of results occur in exchanges of information and views:	Steps 1 to 6 have been followed and reported on for this and many other varieties of English and other languages.

Continued...

8. <i>Inference</i> takes what we know, allowing us to form hypotheses that predict new outcomes.	One possible inference is that speakers of this variety will use sentences like, “I be goin,” with different meaning and in very different contexts than (1) to (3).
9. <i>Applications</i> of science are to continue theory building or to use it for human purposes.	Primary application of this insight into English dialects is to education. Forensic applications rely on variation analysis in voice and authorship identification.

**Table 3.5: “The science of linguistics”, from: McMenamin (2002, Section 2.2.1) (Contents SIC)**

Table 3.5 depicts the scientific aspects that are used in a linguistic study. The table is ordered sequentially from one to nine. Each step depends on the information in the previous one. This study flows in a very similar manner, in that the facts are observed (i.e. the problem statement), then relationships are identified according to how these facts interact with other facts and finally, conclusions are drawn. Point (1) regarding linguistic facts is of great importance to a forensic linguistic investigator, as the investigator can draw upon the observations of human behaviour to create generalisations regarding the author.

### **3.3 Linguistic Variation**

The concept “linguistic variation” is of paramount importance in forensic linguistic investigations. This is the process by which it becomes possible to identify idiosyncratic forms of language use. McMenamin (2002) notes that forensic specialists have to be able to go beyond the constant similarities of evidence, and focus on the variation that allows one piece of evidence to be differentiated from another. He goes on to state that the language used by a particular writer can be attributed to the language used by particular groups or individuals.

McMenamin (2002) states that the first step to determine variation is to observe the data, then identify it as being systematic and then discover the significance of each variant (which in this case would be a writer’s unique style). McMenamin (2002) also explains that language change can only be observed by comparing two synchronic descriptions from two widely separate points in time: this process is known as diachronic variation. He adds that language change is always evident and that language users are aware of it. The changes that one can observe over a short period of time are known as linguistic variation. McMenamin (2002) notes that it is important to possess a solid understanding of linguistic variation in order to be able to

perform authorship identification. The following points help to better understand the existing forms of linguistic variation (McMenamin, 2002).

### 3.3.1 Language Variance as markers of Social Class or of Individual Language Use

There are three types of evidence that can be used to determine individuation (McMenamin cites Inman and Rudin, 1997). These are known as associations, identification of class features, and individuation.

#### 3.3.1.1 Associations

Association concerns itself with the location of similar traits. These traits that are left behind when a crime is committed can be compared to those that have been referenced. Should these traits be found to be similar, an association can then be created between the evidence and the data in the reference, and the strength of the similarity can then be measured. McMenamin (2002) adds that physical evidence by nature is classed as circumstantial, as it does not directly create an association but instead allows for a deductive inference about a possible association.

#### 3.3.1.2 Identification of Class Features

Identification of unique associations allows the forensic auditor to eliminate other possible suspects from the pool according to the suspect's "class characteristics". The class characteristics can be determined by the social influences on the suspect's language usage and comprise details that signify race, social class and gender, for instance.

#### 3.3.1.3 Individuation

According to McMenamin (2002), individuation in respect of linguistics refers to the features of language usage that are formed during a user's language development and re-occur during that person's language usage.

QUESTIONED WRITING	KNOWN WRITINGS
<b>diffulgies</b>	<b>diffulgities</b>
	<i>difulgity</i>
	<i>Defulgity</i>

**Table 3.6: “An example of a possibly unique writing style”, from: McMnamin (2002, Section 3.4.1)**

Table 3.6 depicts a case where the text in question contains the word “diffulties”. Three pieces of evidence are taken from three different sources that are marked as known writings, which can be attributed to a particular author. These known writings are the samples of text that can be verified as being produced by a certain individual. The common characteristic in this table is that the characters “ulgit” occur in all three of the known writings. The corresponding writing in the questioned piece is that of “ulgt”. This proves that the writing styles do not match and that the author of the known writings did not write the text in question. This shows that authorship can be determined by ascertaining a person’s unique writing characteristics.

### **3.3.2 The terms “Language” and “Dialect”**

According to McMnamin (2002), the word “language” refers to the communication of a group of people in general (as in “the Japanese language”), whereas “dialect” describes the communication of a subgroup of those language users who are in some way separated from other speakers either geographically or socially. Geographical separation occurs when long distances are placed between speakers due to barriers, for example, oceans and mountains. Social separation is the result of differing social classes; these classes can be determined by factors such as age, race, income, occupation and surname. This variation assists the forensic linguist to determine the type of class and community to which a speaker belongs.

### **3.3.3 The term “Idiolect”**

According to McMnamin (2002), the term “idiolect” describes a speaker’s personal dialect. He adds that every person will possess a unique grammar usage style, and “idiolect” can be described as a person’s unique linguistic knowledge. McMnamin explains that the goal of linguistic analysis is usually to seek out group characteristics regarding language users, for example, race, gender and social class. The concept of idiolect, in contrast, is concerned with what makes one language user unique from another. It is this concept that this study will centre around, as the use of forensic linguistic analysis for authorship purposes is concerned solely with the individual.

### 3.4 Linguistic Variation Analysis Techniques

McMenamin (2002) notes that two methods of variation analysis exist. These are known as the bottom-up model and the top-down approach. The bottom-up model aims to identify patterns and forms in the text in order to create various hypotheses regarding the writer's writing style. The top-down system searches for a predetermined set of stylistic features that would enable the investigator to come to conclusions regarding the characteristics of writers from certain speech communities. It also is important to note that language is a combinatory system that is rule-governed and systematic in nature; therefore sound can be grouped into words, words then grouped into phrases, and phrases into sentences, and finally, sentences into discourse (McMenamin, 2002).

#### 3.4.1 Linguistic Variables

Linguistic variation refers to the ability of a language user to say the same thing in more than one way, for example, "can't, cannot" and "can not". Linguistic variation is most commonly used with two or more variants. These variants are represented by a linguistic variable that shows the degree and type of variation for a particular type of speech usage (McMenamin, 2002). The information can then allow the investigator to compare the frequencies of the known writing with the frequencies of the questioned ones.

Five Variables with Two Variants Each	<u>Known</u> Letters		<u>Questioned</u> Letter	
	n	%	n	%
<b>1. Verb to be</b> PRESENT: I know you <b>are</b> a good man. ABSENT: Know you good man ....	54/69	78%	3/12	25%
	15/69	22%	9/12	75%
<b>2. Auxiliary Verb do</b> PRESENT: But <b>do</b> not think he will take .... ABSENT: He say he not want my money.	17/17	100%	3/13	23%
	0/17	0%	10/13	77%
<b>3. Other Auxiliary Verbs</b> PRESENT: I <b>was</b> losing hope then they come .... ABSENT: You tell me I taking to much dope.	42/50	84%	11/28	39%
	8/50	16%	17/28	61%
<b>4. Article a</b> PRESENT: So I know you are <b>a</b> good man. ABSENT: Know you good man and will do....	19/58	33%	0/34	0%
	39/58	67%	34/34	100%
<b>5. Past Tense Marker</b> PRESENT: The nurse <b>told</b> me that Ed asked .... ABSENT: Nurse <b>tell</b> me to put ....	188/209	90%	21/95	22%
	21/209	10%	74/95	78%

**Table 3.7: “Relative proportions of variants for each of the five variables”, from: McMenamin (2002, Section 3.5.2)**

Table 3.7 above indicates the system used to depict the percentages of present and absent category variables. The known letters column refers to the sample of writing that is verified as being that of the writer in question, whereas the questioned letters column is the one which the investigator wishes to determine authorship.

### 3.4.2 Sentence Variation

Language is said to be a system that is rule-governed (McMenamin, 2002; section 3.4). Every language user has a diverse number of word choices when constructing sentences.

1	2	3	4	5	6	7
This The That Our	pitcher goalie flautist contestant survivor captain	is was	probably certainly surely I think I believe we think we believe one hopes	a another	gorgeous pretty smart lovely beautiful ravishing wonderful striking	one. girl. woman. lady. female. player.

**Table 3.8: “Sentence string with seven opportunities for word choice”, from: McMenamin (2002, Section 3.6)**

Table 3.8 above depicts a sentence comprising seven words with various possible word types that could be used for each word.

Sentence #1:

The	pitcher	is	probably	a	gorgeous	one.
-----	---------	----	----------	---	----------	------

....

Sentence #36,864:

Our	captain	was	one hopes	another	striking	player.
-----	---------	-----	-----------	---------	----------	---------

**Table 3.9: “Two of the 36,864 possible choices”, from: McMenamin (2002, Section 3.6)**

Table 3.9 is simply an example of two of the possible sentence outcomes from table 3.8 above. As we can see, there exists a total number of 36,864 possible combinations that can be made. This illustrates the issue of how a language user can

utilise numerous possibilities and combinations of words in everyday verbal and written language usage.

### **3.5 Relationships between Forensic Linguistics and Other Elements**

This section aims to show the indirect relationships that exist between forensic linguistics and various other factors.

#### **3.5.1 Relationship between DNA and Language Stylistics**

According to McMenamain (2002), a relationship exists between DNA structures used for identifying a person, and grammar usage for identifying a language user. He notes that comparison of DNA to language can be done in two predominant ways. Firstly, is the language used to describe the respective systems and secondly, the comparison of the various elements and categorisation of systems. McMenamain (2002) adds that it only takes the smallest percentage of DNA variation to differentiate people (0.5%). Language depicts similar characteristics, in that it is commonly used among many people, yet it is only the characteristics that are left over from the common pool of language use that reflects the idiolect of a particular user.

McMenamin (2002) states that in order to measure language in a similar system to DNA, one must follow the given system.

- i.** A corpus of writings must be acquired
- ii.** A database of style markers must then be created
- iii.** Markers can then be said to occur within a generalised frequency in specific writing contexts. Dependence or independence of these markers on each other will also have to be determined at this stage

Once these factors have been processed, the questioned and known writings will be compared and their similarities and differences noted. Writers whose idiolects do not match the frequencies of the class features are excluded from the pool. Individuation is then established by examining multiple markers to determine the lowest frequency possible for the remaining suspects. As more markers are analysed, the population of suspected writers drops.



### **3.5.2 Relationship between Linguistics and Document Examination**

Document examination deals with the physical traces that can be used to find characteristics of the document. For example, QD (questioned document) analysis can be used to find the pen type, ink, writing surface, computer and similar aspects. With regard to handwriting, the QD examiner will look at features such as letter size and slant, spacing and line quality, amongst others. Typing features such as font style, spacing between letters and the type of machine, can be examined (McMenamin, 2002).

There is a difference between linguistic and QD examination (McMenamin cites Huber and Headrick, 1999: 79). Huber and Headrick add that the difference can be determined by the significant lack of style characteristics analysis in QD examination. QD analysis is predominately concerned with the analysis of aspects such as punctuation and abbreviations rather than word usage.

### **3.5.3 Relationship between Linguistics and Software Forensics**

Stylistic analysis for software is a new addition to the realm of authorship identification. Two programs that have exactly the same function may have very different coding (McMenamin, 2002). Stylistic analysts have determined a different set of markers in order to enable the determining of authorship for software. These are, amongst others variable names, number of global and local variables, case usage, debugging comments and line length. Software programs that allow for the detection of plagiarism in software are increasingly being used at academic institutions. McMenamin (2002) adds that a program known as MOSS (Measure of Software Similarity) created by Alex Aiken, was released as freeware to many professors in 1997.

## **3.6 Stylistics and Forensics**

McMenamin (2002) characterises style as the selection of particular attributes of a variable with regard to human behaviour. Once developed, a style of doing something remains with a person. With regard to language, style describes the characteristic linguistic variation in a person's spoken or written language. As indicated before, this variation is dependent on a number of factors such as age, gender, race and education. With regard to written language, style refers to the ways in which language is used by individuals, or during particular periods or in specific

genres. Writing style is the recurrent choices the writer makes from his/her subconscious habits. Identifiable markers here will fall into two categories: variation within a norm or deviation from a norm. Variation within a norm refers to grammatically correct forms of language usage, while variation from a norm refers to grammatical usage that could be described as incorrect.

Norms can differ according to various factors, which is why it is important to correctly define a norm before it can be used as a standard for identifying variation. Norms can be described in both linguistic and statistical fields. In the linguistic category, there exist two types of “norms”. These are known as prescriptive and descriptive norms. Descriptive norms reflect what the writer perceives as being grammatically acceptable, whereas prescriptive norms refer to the social sense of what language usage is considered correct and acceptable. Statistical norms describe linguistic norms in a frequency distribution (McMenamin, 2002).

### **3.6.1 Linguistic Stylistics**

Linguistic stylistics with regard to a single language user is said to be the scientific analysis of individual style markers derived from the idiolect of that person. With regard to a group of language users, it can be described as the analysis of class style markers of that particular group. Linguistic variation makes use of a “norm” to determine if variation is within a norm, or deviates from it. The “norm” mentioned here is determined by speech pattern variation in the particular speech community in which authorship is being determined (McMenamin, 2002). There will be as many norms as there are groups and there are as many sets of norms as there are dialects in a language.

TYPE OF NORM	THE NORM	VARIATION WITHIN THE NORM	DEVIATION FROM THE NORM
<b>PRESCRIPTIVE</b>	<i>Examples</i>	<i>Examples</i>	<i>Examples</i>
Grammatically correct	I <i>am</i> going now.	I'm going now.	I <i>be</i> goin' now.
Socially appropriate	I'm afraid you're too late.	Sorry, the shop is closed.	Get the hell out of here!
<b>DESCRIPTIVE</b>			
Prestige: U.S. standard	We have enough money.	We've <i>got</i> enough money.	We <i>gots</i> enough money.
Choice of variety: teenage	<i>Hey, man!</i>	<i>Hey, dude!</i>	<i>Hello, sir.</i>
Class: age	That's a <i>cool</i> idea.	That's a <i>neat</i> idea.	That's a <i>swell</i> idea.
Regional: U.S. dialects	"a quarter <i>to</i> eleven"	is <i>before</i> - / <i>of</i> - / <i>till</i> - eleven	"eleven <i>less</i> fifteen"
Situational: at work	"Where's the <i>restroom</i> ?"	"Where's the <i>bathroom</i> ?"	"Where can I <i>take a leak</i> ?"
<b>QUANTITATIVE</b>			
How often forms are used	We <i>are</i> here. (e.g., 10%)	We're here. (e.g., 80%)	We here. (e.g., 10%)
In a defined social context	It <i>is</i> me. / It's <i>me</i> . (85%)	It <i>is</i> I. (10%)	It <i>be</i> me. (5%)

**Table 3.10: "Examples of Linguistic Norms", from: McMenamin (2002, Section 6.2)**

Table 3.10 above shows three of the many categories of norms that exist (prescriptive, descriptive and quantitative), as well as how the associated phrases can vary within and deviate from it.

### 3.6.2 Analysis Techniques

McMenamin (2002) cites Wachal's three methods of authorship analysis: resemblance, consistency and population. The resemblance model is the best known model for authorship analysis and is most commonly used when the pool of suspects is narrowed down to one or a small number of writers. The consistency model is used when the investigator needs to determine if two or more oral or written texts were authored by the same person. Lastly, the population model is used when one cannot narrow down the sample set to just one or two writers. These models can be used in combination with each other.

### 3.6.2.1 Software Tools for establishing Concordance

According to the Kings College of London website, henceforth referred to as Kings College of London (2007), concordance allows the investigator to locate every occurrence of a word or set of words within a corpus, for instance, any electronic document subject to forensic auditing. This will allow for detection of various patterns of word usage, thereby allowing the investigator to base an argument on the patterns as evidence. Concordance analysis is particularly useful in determining the intended meaning of a certain set of words, as it is often the case that a set of words can be used in such a way as to infer meaning very differently to the particular words taken on face value.

Kings College of London (2007) adds that in the past, before computers were available, concordance analysis was a very tedious and time-consuming process. Words had to be manually located within books, journals and other repositories. However, with computers, this task can now be done very quickly and efficiently.

Several computer packages exist for facilitating concordance analysis. Below, a few of the downloadable computer-based software concordance programs are discussed briefly as examples of forensic analysis tools available for document analysis.

#### 3.6.2.1.1 Phrase Context

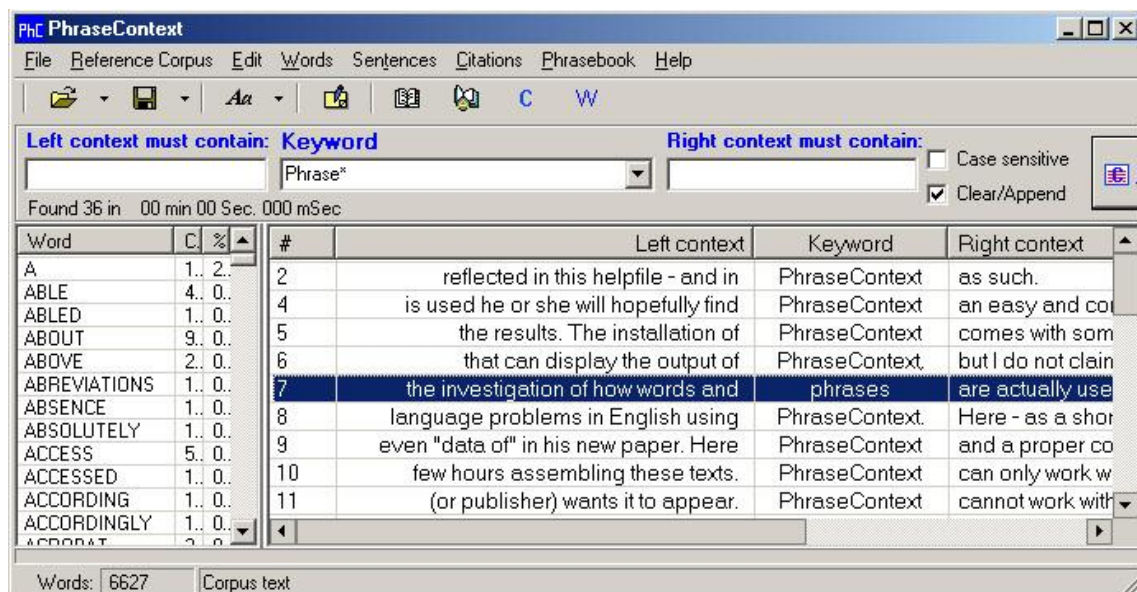


Figure 3.4: Sample Image of the Phrase Context Program GUI, from: Mortensen (No Date)

This program is one of the concordance analysis programs that can be used by investigators to assist in determining authorship. The following facts and points are closely summarised from Mortensen (No Date), which is referenced in the bibliography.

**i. Features**

- PhraseBooks can be shared across a network with other academics and can be exported in XML format
- The PhraseBook is a fully editable XML-file
- No database engine is needed
- Wordlists can be calculated for large corpora
- Keyword in context can be quickly created for a series of words
- Statistical analysis functions include factors such as the T-score and Z-score
- Collocations can be derived and outputted
- Sentence length analysis can be outputted as a chart in .bmp format
- Citations and wordlist can be saved for quick reloading
- The calculation of lexical density, depicting the overuse or underuse of certain words, is allowed
- Single or multiple file corpus can be used
- It contains a corpus viewer for interactive work

These are not all the claimed functions of this software package.

**3.6.2.1.2 Concordance**

The following facts and points are closely summarised from Concordance (2009), which is referenced in the bibliography.

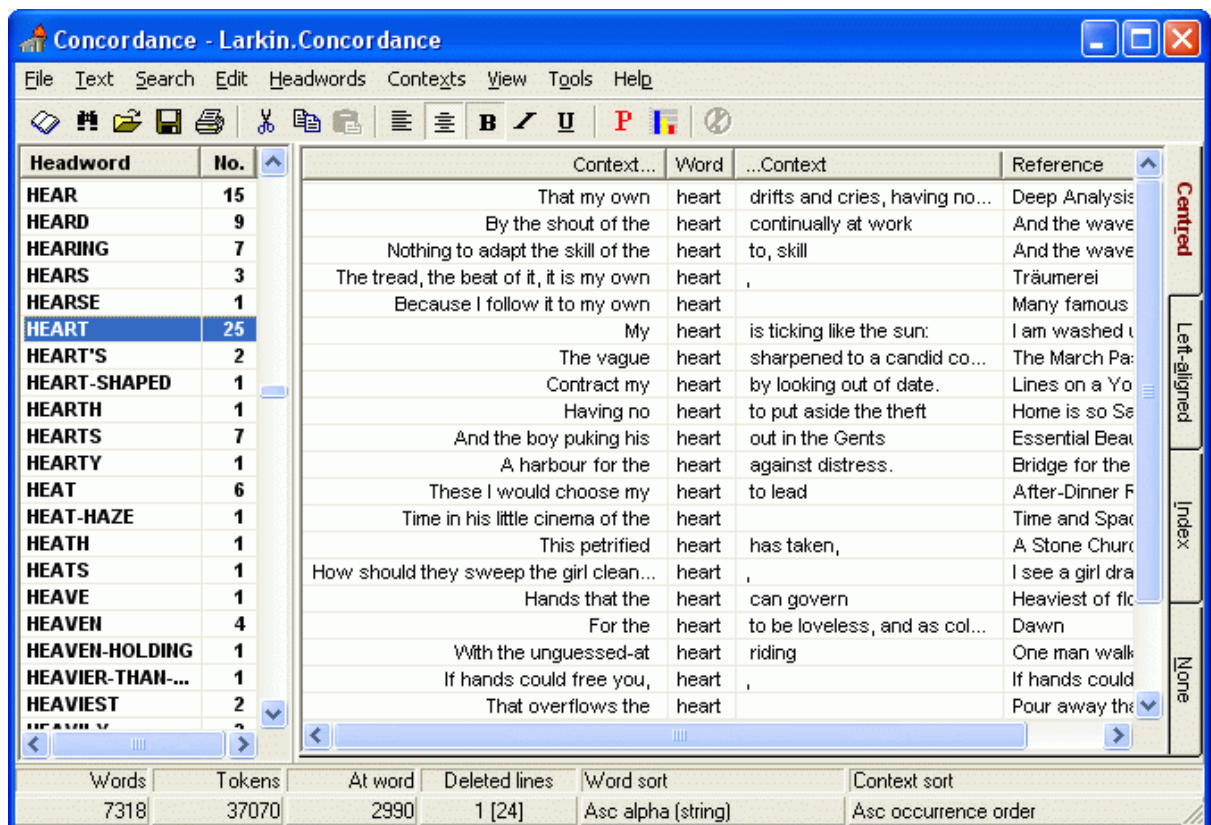


Figure 3.5: Sample Image of the Concordance Program GUI, from: Concordance (2009)

### i. Functions

The program can output indexes and word lists, perform word frequency counts, draw up comparisons of different word usages, perform keyword analysis, search for phrases and publish data to the Internet.

#### 3.6.2.1.2 MonoConc Pro and Paraconc

These are two software programs for concordance analysis. MonoConc Pro has the following facts and points that are closely summarised from Athelstan (2007), which is referenced in the bibliography.

### i. Corpus Analysis features

MonoConc can accept any alphabetic language and has no limit on the size of the corpus it can analyse. Once a corpus has been uploaded, it can be edited at any time.

## ii. Searching

Searches can be run for either words or phrases; advanced search options include tag searching or regular expression searches. Wildcard characters can be accepted and batch searches can be run with multiple differing search criteria.

## iii. Frequency information

Frequency information can be gathered on the search words and corpus frequency can be determined as well as batch frequency. The batch frequency system allows for the determination of frequencies over a batch of texts at once. This system can greatly speed up the process of calculating word frequencies over a large volume of texts.

## iv. GUI

The program can display data in Key Word In Context (KWIC) or sentence format and contains a large context window. A graph of hits is also displayed. The KWIC concordance system is simply a term used to describe concordance systems such as that in figure 3.5, where the keyword is displayed in the centre of the page with the adjacent text appearing to the left and right of it.

ParaConc is simply a concordance program that can perform concordance analyses on different languages. It has a very similar graphical user interface to its counterpart, MonoConc.

### 3.6.2.2 Software Tools for Word Frequency Analysis within Corpora

The following section describes some of the software packages available for performing word frequency analysis on corpora, referring to a few of the available packages and their stated features.

#### 3.6.2.2.1 T-Lab



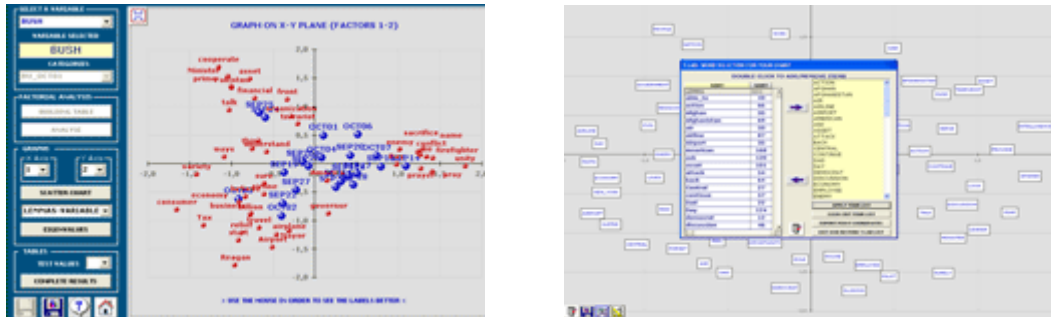


Figure 3.6: Sample screens of the T-LAB program GUI, from: TLAB (2009)

T-LAB has been designed to run in a windows environment. It outputs tables and graphs that are simple to interpret. This package can automatically perform many other statistical analysis techniques on the inputted text. Automatic lemmatisation is also available for several languages. Texts can be inputted in compatible ASCII formatting.

### 3.6.2.2.2 Weka

Weka is a feature packed data mining software solution. A major aspect of Weka is the knowledge explorer described below, adopted from the website of the University of Waikato (2009), hereon refer to as Weka (2009).

#### i. The Weka Knowledge Explorer

The Weka Knowledge Explorer is the focal point for the majority of the main functions of this package. The explorer (figure 3.7) contains the following items: filter, classifier, cluster, association, select attributes and visualise. The visualise tool allows for a two-dimensional description of data sets and classifiers.

#### ii. Pre-process Panel

The initial phase of the knowledge exploration is known as the pre-process panel. Here one can input data and browse attributes (see figure 3.7).



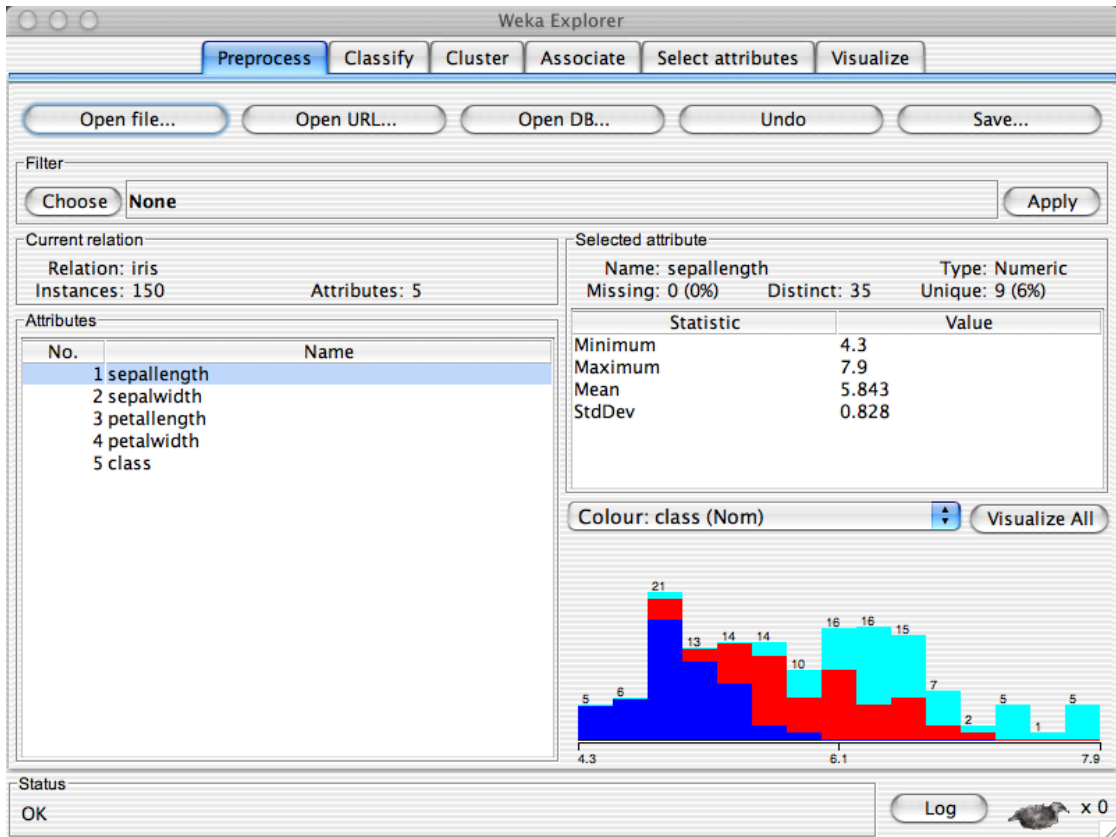


Figure 3.7: Pre-process Panel, from: WEKA (2009)

### iii. Classifier Panel

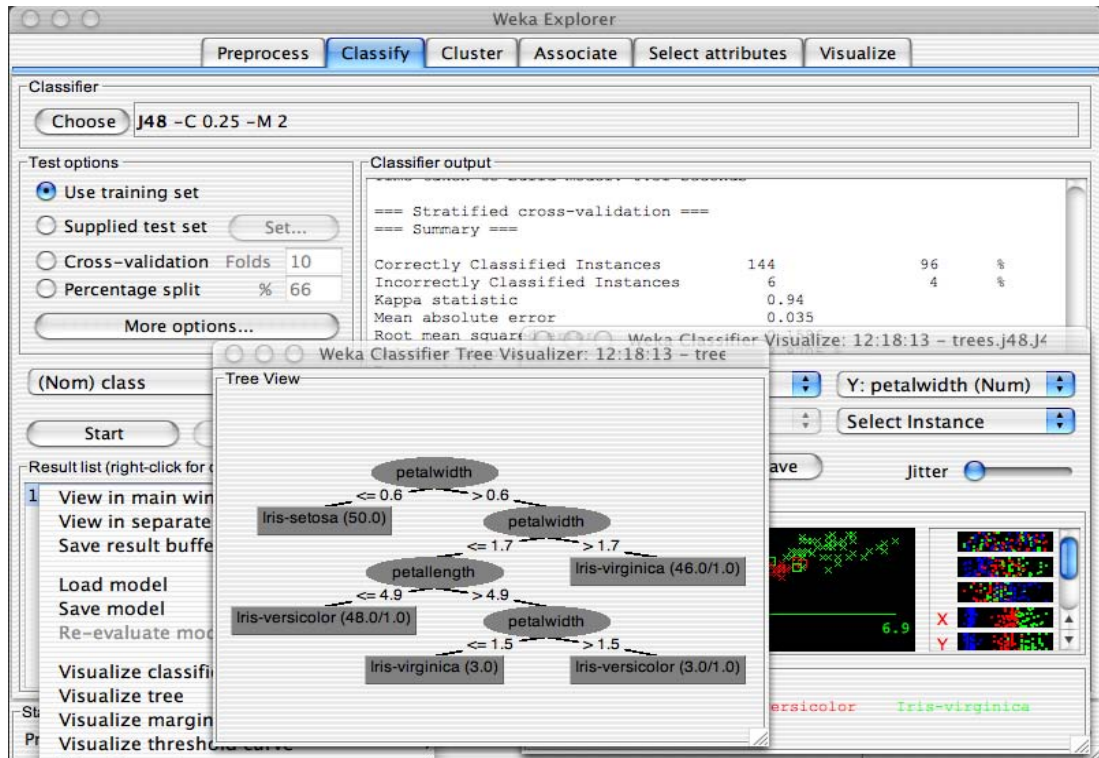


Figure 3.8: Classifier Panel, from: WEKA (2009)

This tab (figure 3.8) allows for the classifiers to be executed on the inputted data. Any errors can be depicted in a pop-up tool, which can also display a decision tree if the program generated one.

#### iv. Cluster Panel

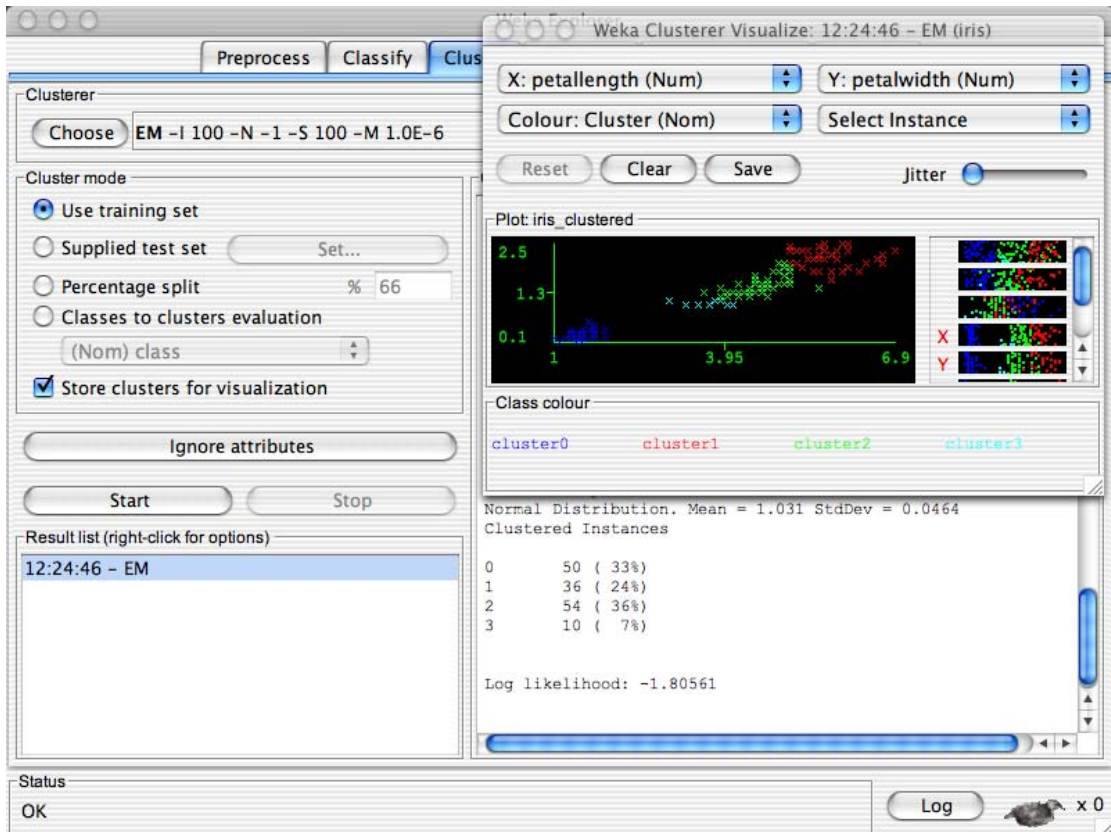


Figure 3.9: Pre-process Panel, from: WEKA (2009)

This tab (figure 3.9) simply allows for the clusterers to be executed on the data, the output from which can be displayed on a pop-up tool. This is one of the various statistical functions available within WEKA.

## v. Associate Panel

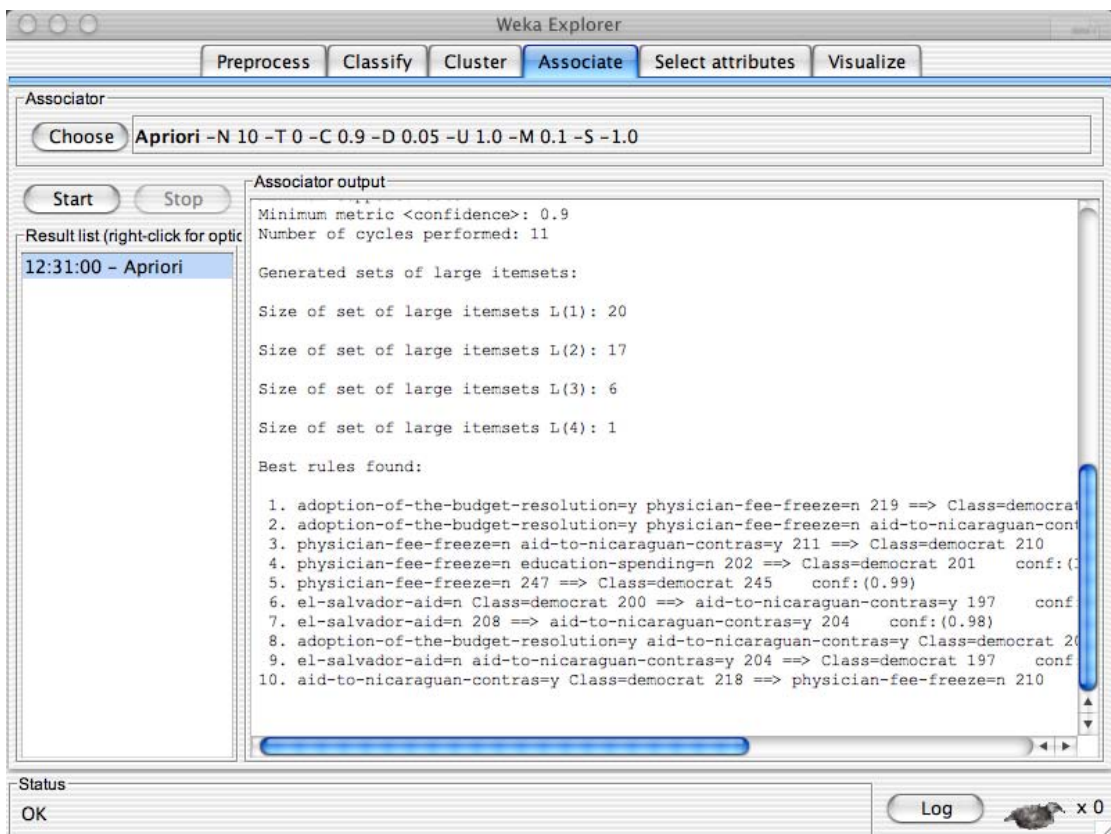


Figure 3.10: Associate Panel, from: WEKA (2009)

This tab (figure 3.10) is used specifically for data mining purposes. Data mining is the process of sifting through often large quantities of data using constraints in order to acquire specific data. Within WEKA, one simply chooses the required associator and clicks *start* to begin the data mining process.

## vi. Select Attributes Panel

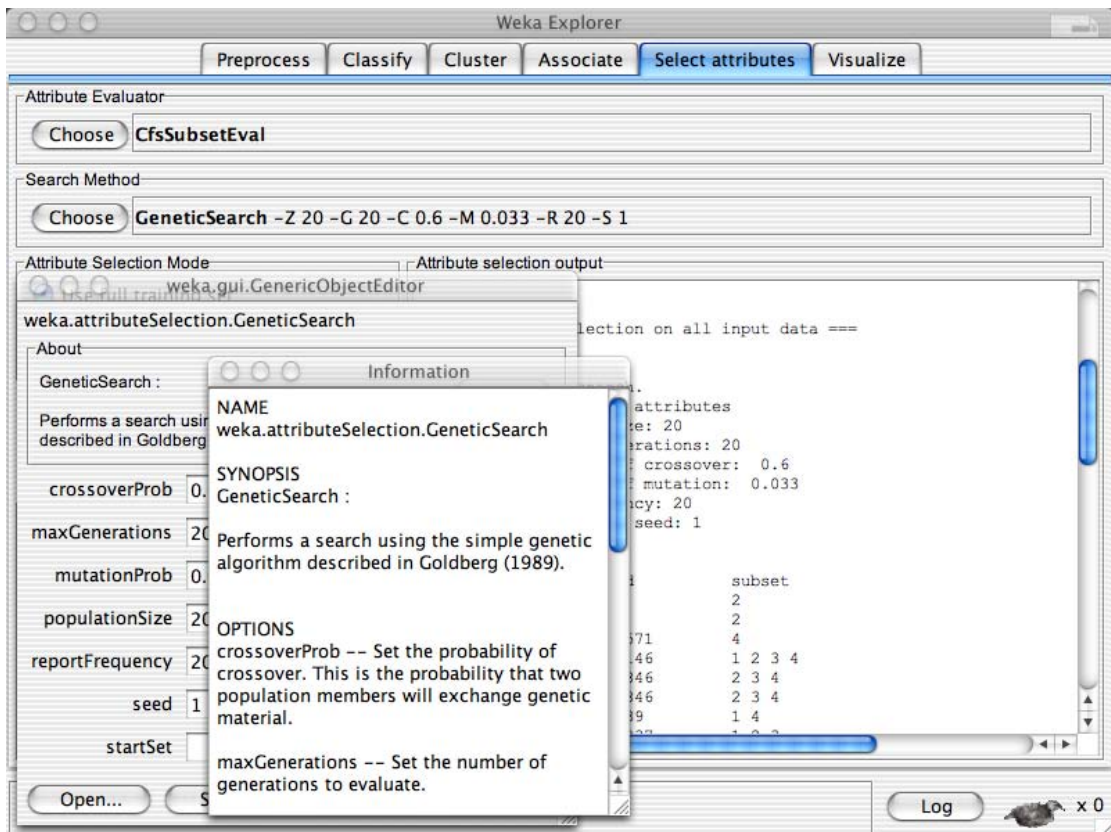


Figure 3.11: Select Attributes Panel, from: WEKA (2009)

The select attributes tab (figure 3.11) allows the user to select the most important aspects of the dataset. If the selectors transform an attribute, then the corresponding data can be displayed in the pop-up tool.

## vii. Visualize Panel

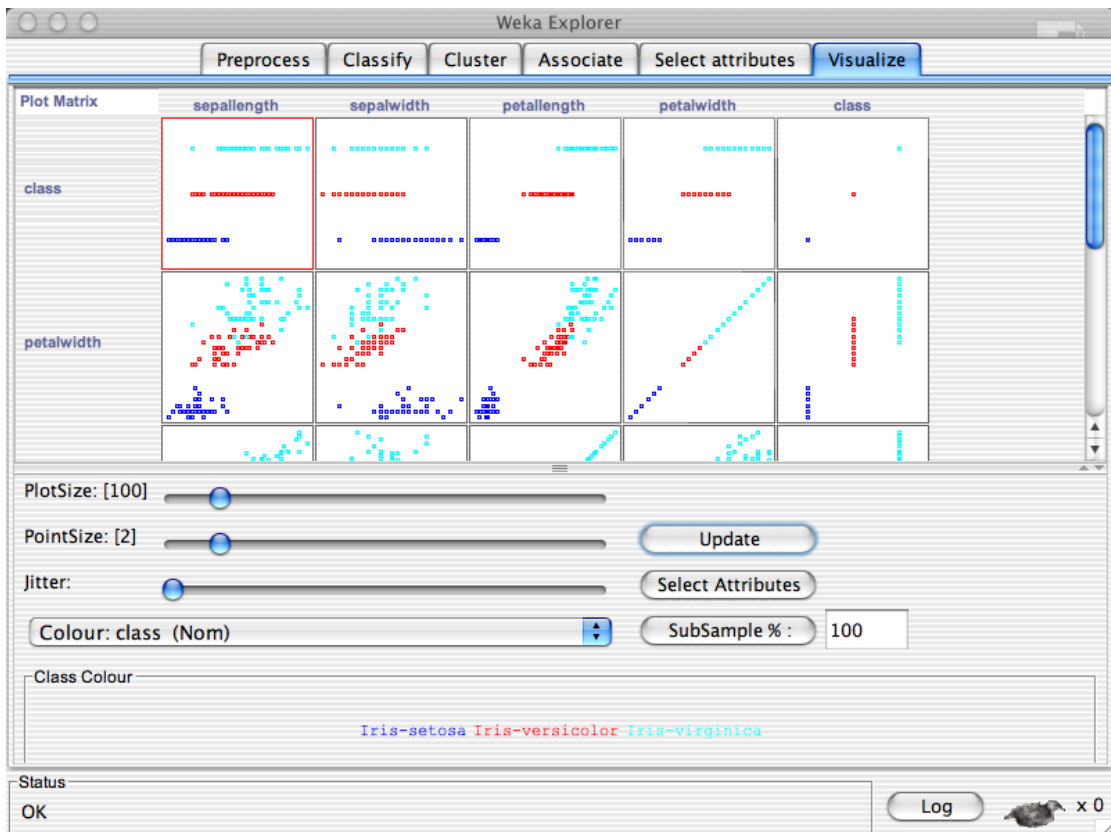


Figure 3.12: Visualize Panel, from: WEKA (2009)

The visualise panel (figure 3.12) above uses a scatter graph to display information for the dataset. All aspects of the graph can be adjusted and the attributes to be displayed can be selected. Attributes can be viewed in one or two dimensions, and attribute bars provide a summary of the power of the attributes.

### viii. Interactive decision tree construction

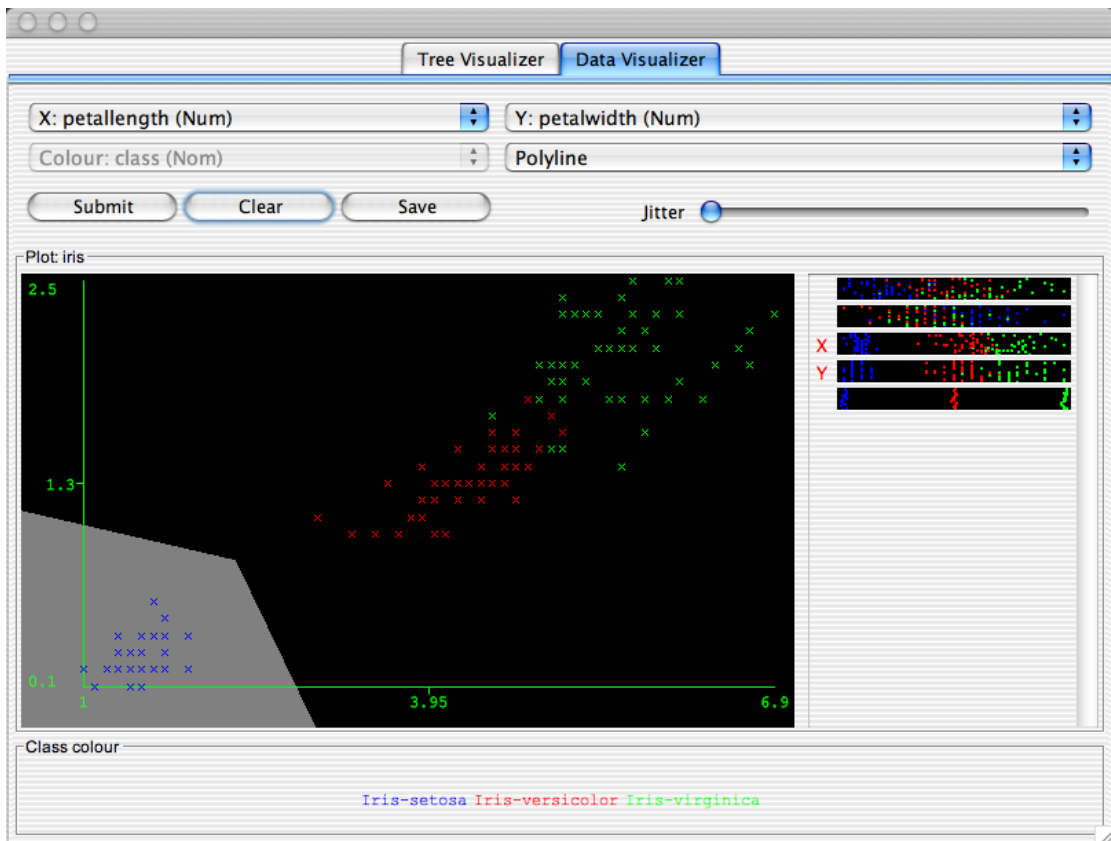


Figure 3.13: Interactive decision tree construction Panel, from: WEKA (2009)

The program allows for the creation of an interactive decision tree. The aspects of this tree can be edited and manipulated at any time in the construction phase.

## ix. Neural Network GUI

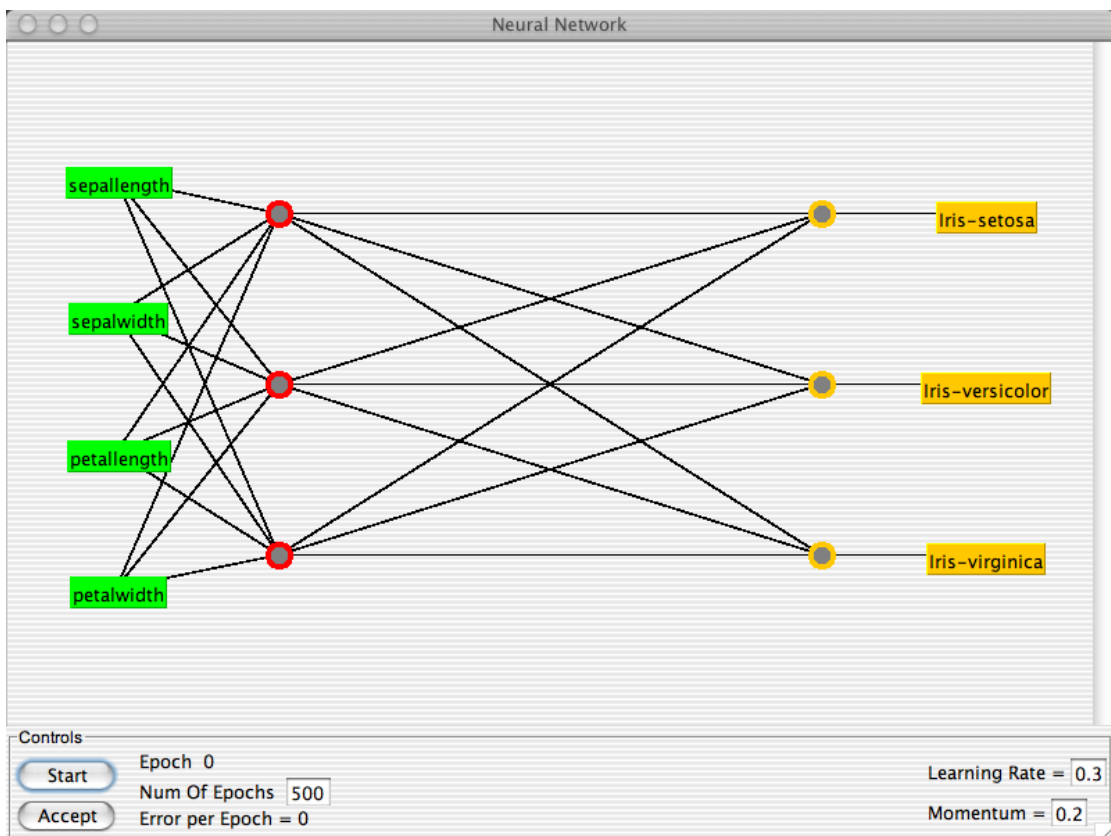


Figure 3.14: Neural Network GUI, from: WEKA (2009)

Figure 3.14 depicts the Weka neural network GUI. This shows the multilayer perceptron and its parameters, which is yet another statistical function of this package.

### 3.6.2.2.3 Wmatrix

The following facts and points are closely summarised from Rayson (2008), which is referenced in the bibliography.

This package is one of the various programs that are in existence for corpus analysis. Some of its features include a web interface and corpus analysis methods such as concordance and frequency analysis. The tool can be run via an Internet browser and is compatible with many different types. It is also a multiplatform package that runs on Windows, Mac Unix and Linux operating systems.

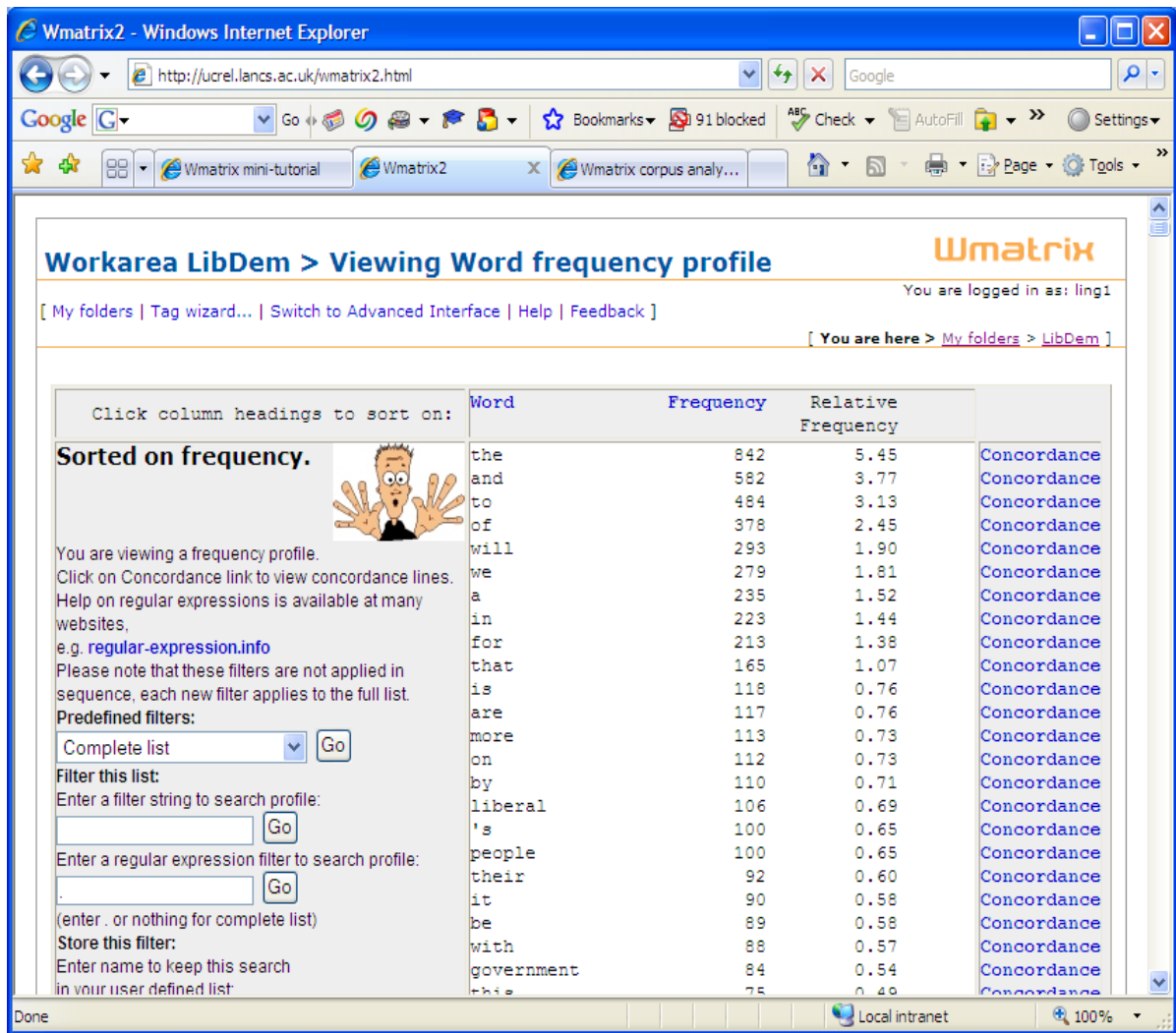


Figure 3.15: Wmatrix GUI, from: Rayson (2008)

Figure 3.15 above depicts the Wmatrix word frequency distribution output screen. The words are displayed on the left with the corresponding frequency on the right, followed by the relative frequency. The list can be sorted by various attributes, frequency being one. This makes it easier to view large sets of frequencies.



## i. The Tag Wizard

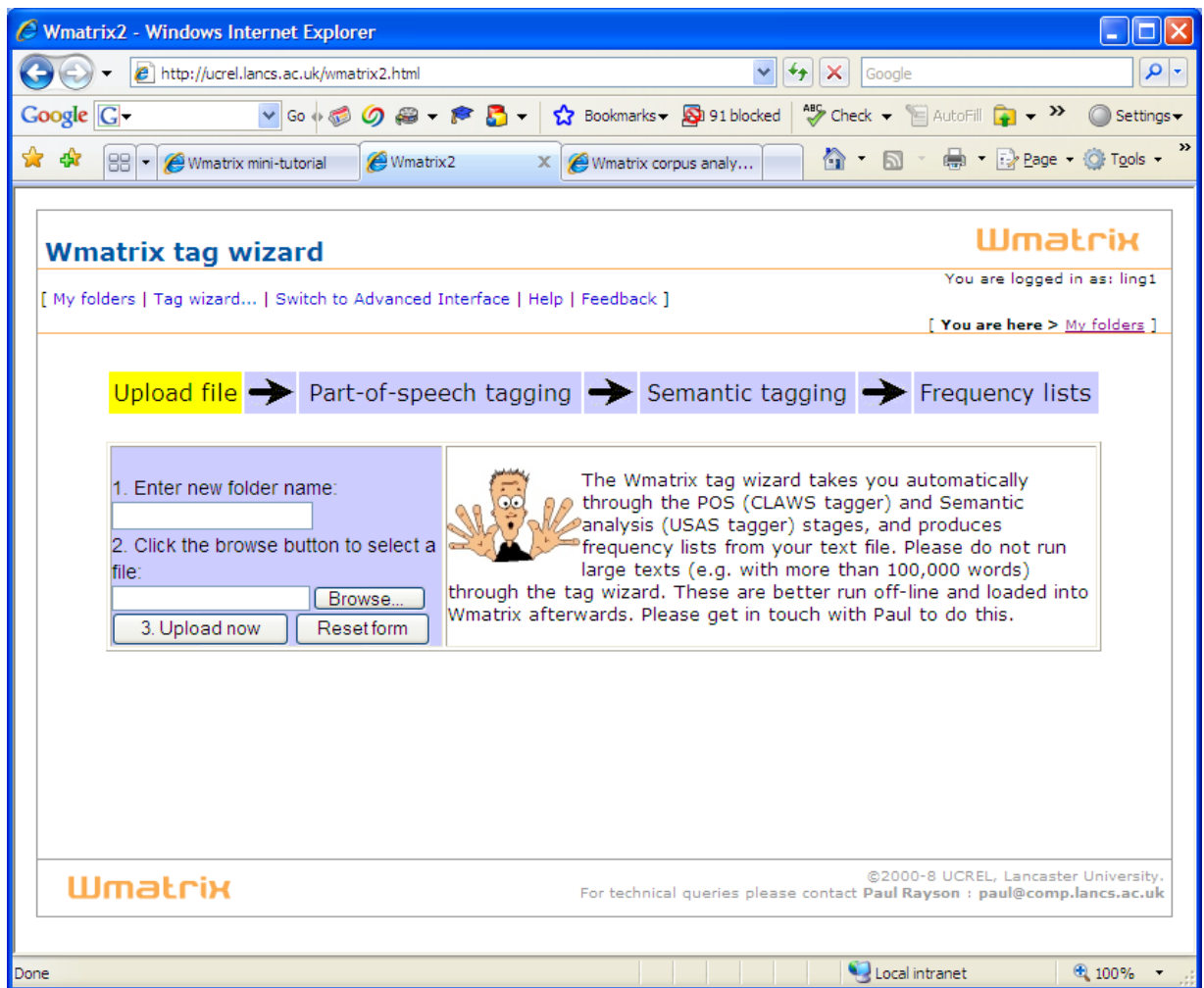


Figure 3.16: Wmatrix Tag Wizard, from: Rayson (2008)

Figure 3.16 depicts the tag wizard screen. This screen is for the uploading of files for tagging purposes.

## ii. Viewing folders

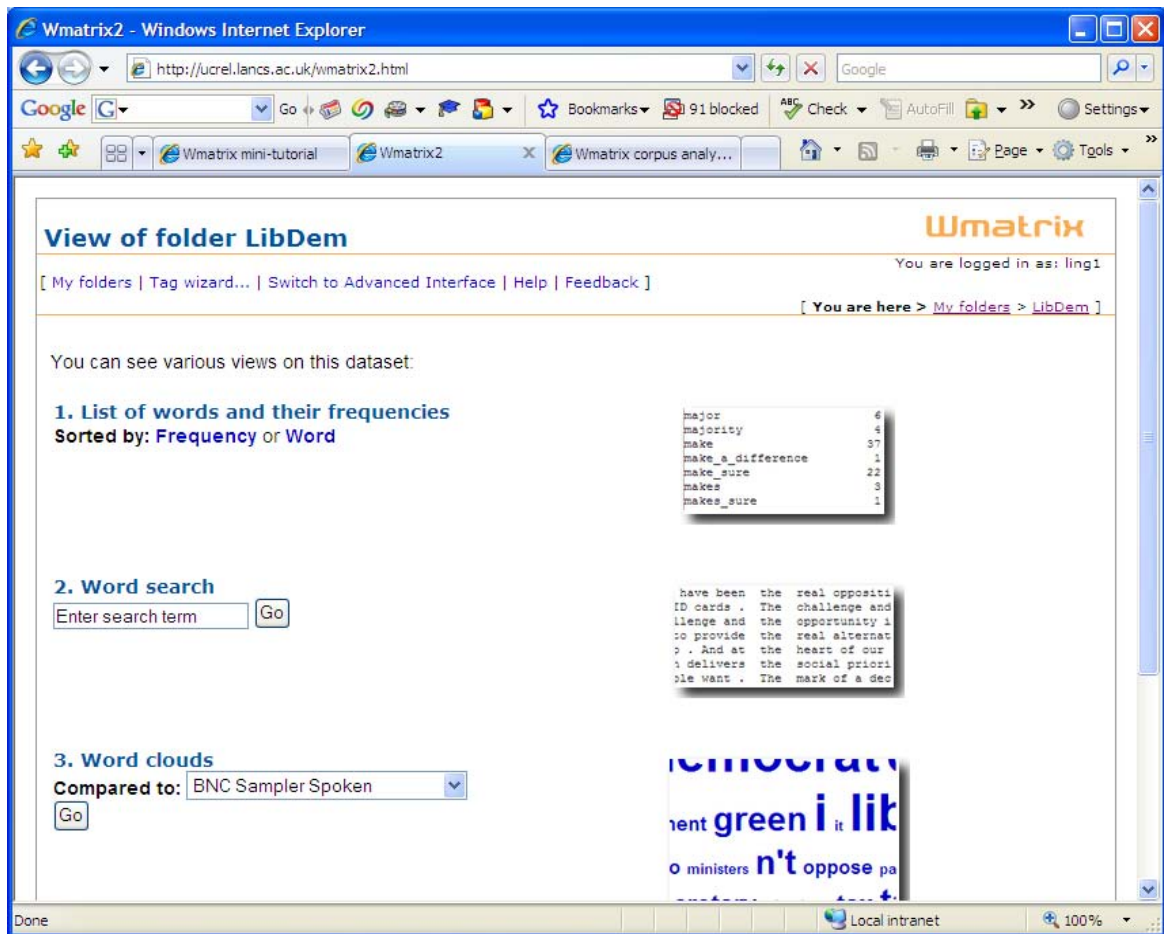


Figure 3.17: Wmatrix View Folder Function, from: Rayson (2008)

Figure 3.17 above depicts the view folder screen, where the user can view various aspects of a created folder e.g. the contents and frequency profiles.

## iii. Interfaces

The user can choose between the simple and advanced interface options.

## iv. Frequencies

The folder view screen allows the user access to the frequency list of their corpus. These lists can be sorted either alphabetically or by frequency.

## v. Concordances

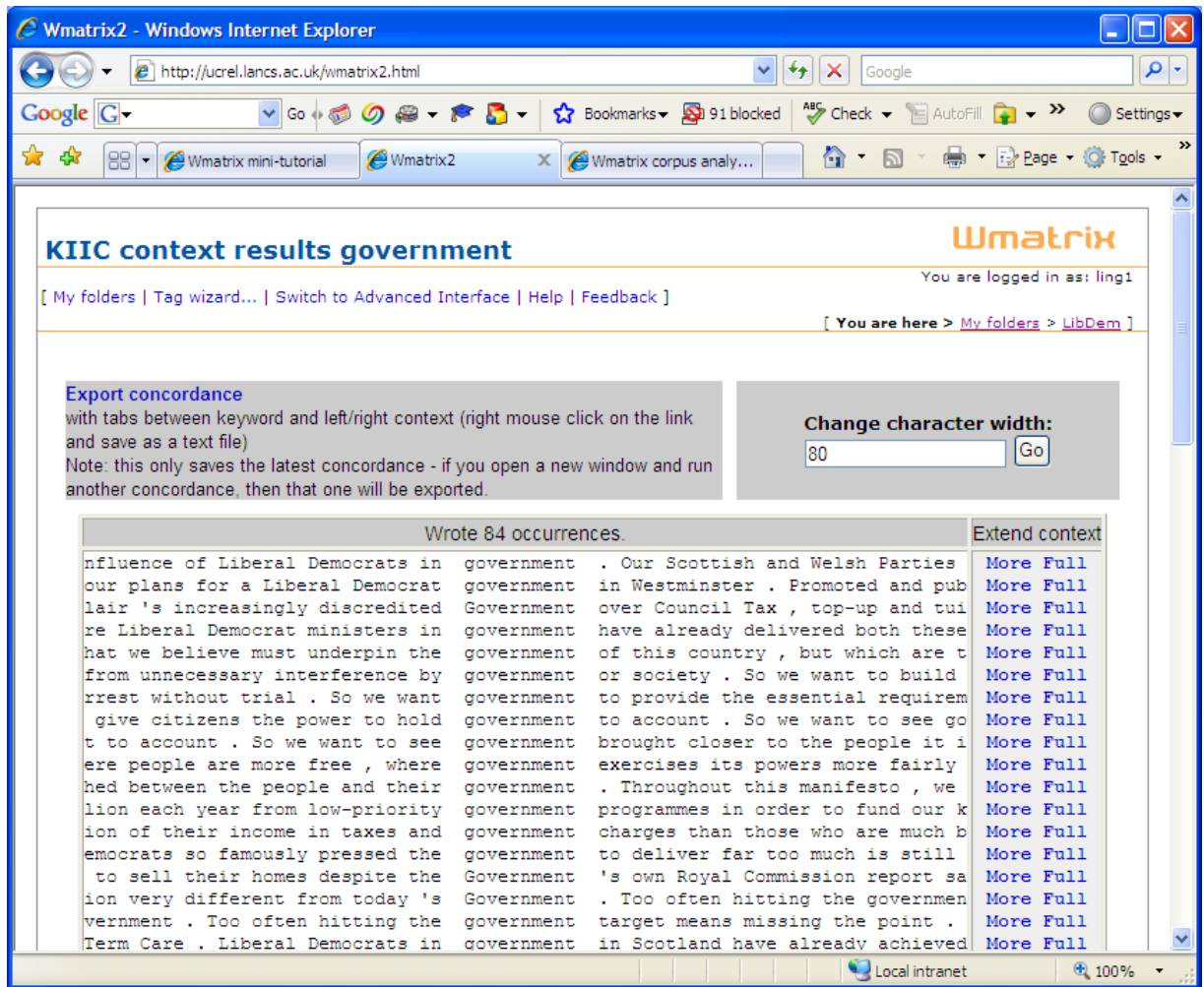


Figure 3.18: Wmatrix Concordance Function, from: Rayson (2008)

Figure 3.18 depicts the concordance function of Wmatrix. This screen shows the concordances for the corpus using the word "government". A handy feature of this package is that should the user require more detail as to a particular occurrence, the user can simply click either the "More" or "Full" options located in the "Extend content" column.

vi. Key words, key POS and key domains: Comparison of frequency lists



Figure 3.19: Wmatrix Compare Frequency Lists Function, from: Rayson (2008)

Figure 3.19 depicts the compare frequency lists function. Here frequency lists from one corpus can be compared to those of another, and the font size and underlining depict different text characteristics for the word in question. In the screen above, the larger the text size, the more significant it is. The search constraints can be assessed according to various levels that are built into Wmatrix.

### 3.7 Framework for Authorship Identification Analysis

This section aims to give a basic understanding of the type of systematic approach that is required when going about an authorship identification investigation. The following points are from McMnamin (2002).

Terms	Definitions
<b>VARIATION</b>	
Variable	A stylistic feature of the writing, also called <i>characteristic</i> or <i>style marker</i> .
Variation	The combination of all occurrences of the same variable found in a set of writings.
Similarity	A variable that is present in two sets of writings.
Dissimilarity	A variable that is present in one set of writings, but absent in another set.
Nonoccurring variable	A variable with no opportunity to occur in the language used in a set of writings.
<b>INDIVIDUALIZATION</b>	
Class characteristic	A variable that is found in the written dialect or language of a definable group.
Individualizing characteristic	A variable that is not a class characteristic, but not necessarily unique to the writer.
Idiosyncratic characteristic	A single variable that is unique to a given writer. This is infrequent in language.
Range of variation	The unique combination of all occurrences of all variables found in a set of writings.
<b>WRITINGS</b>	
Natural writing	Writing done in the context of its purpose, with little attention to the writing process.
Request writing	Writing done for the purpose of providing a writing sample, often via dictation.
Context of writing	Purpose, intended reader, topic, medium (paper), instrument (pen), time, place, etc.
Comparable writings	Two or more sets of writing that share the same or similar contexts of writing.
Quantity of writing	The amount of writing needed to assess the writer's range of variation.

**Table 3.11: “Definitions related to variation, individualization, and writings”, from: McMnamin (2002, Section 6.4) (Contents SIC)**

Table 3.11 above depicts some of the pertinent terms and their definitions relating to authorship identification. These should be included in any reports presented to clients or the courts, for examination.

### **3.7.1 Organisation of the Case**

Clients must be required to submit a signed mandate giving the investigator authority to complete various tasks. This mandate must additionally contain a clear indication of the authorship problem as they perceive it, a list of attached documents and an agreement to the conditions for expert witness testimony. All documentation should be arranged chronologically according to type. Relevant evidence should be photocopied and these used to work from. All correspondence should be noted and saved for future reference.

### **3.7.2 Understanding the Problem**

The investigator should attempt to understand the problem on her own. Once this is done, she should discuss her understanding of the situation with the client. It should be ensured that the work done falls within the scope of the mandate letter and if asked to go beyond the requirements of the mandate, a new one should be drafted and signed by the client (containing the amended scope). The research questions should be identified and classified for both descriptive and quantitative analysis. It is important to note that the research questions and problems will change and evolve as the investigation proceeds. In this phase, one of the three authorship methods (resemblance, consistency or population) must be chosen.

### **3.7.3 Method of Investigation**

McMenamin (2002), states that the investigator must endeavour to collect all known and questioned texts before the analysis can begin. Then the range of stylistic variation must be analysed. Lists of all variations must be made, including variant and invariant forms. Variant and invariant forms like “may not” for which there is no variant, as opposed to “can not” which has the variance “cannot” and “can’t”, must be acknowledged. This allows for the detection of variation in known writings that may not exist in the questioned ones and *vice versa*. Style markers must be identified using deviations from the norm and variations within the norm. Single occurrences of variation must be noted as well as those that occur more than once, representing a recurring habit. Should the corpora be large, they must be scanned and OCR processed. The resulting text can then be fed into one of the many available corpus analysis software tools.

### 3.7.4 The Format for a Forensic Audit Report

Stylistic variation can be characterised with the terms and definitions provided by McMenemy (2002), in section 3.7 in this thesis. The forensic report should identify style markers at all linguistic levels. It should assess the format (layout of the document under investigation, punctuation, spelling, word formation, syntax, lexical variation, semantic variation, functional variation and interference features from the author's mother tongue, in the case of persons writing in a language other than their mother tongue). Style markers need to be identified for both specific markers and descriptive analysis. Specific style markers must contain the following features: format or layout of the document, punctuation, spelling patterns and mistakes, word formation, syntax (sentence structure and punctuation), lexical variation (the usage of specific words and phrases), functional variation and interference features of other languages in the English corpus. Descriptive analysis refers to the describing of the average of the deviation and variation from the norm for the questioned and known writings.

Quantitative analysis can be used to compare the variation between known and questioned writings. McMenemy (2002) states that the probability of occurrences of variables between the known and questioned materials can be measured quantitatively using frequency analysis. Frequency distributions can also be used to determine high and low frequencies of function words, common words and collection of words amongst others. These results can then be represented by means of various graphs.

Once the texts are thoroughly analysed, the conclusions can be drafted. These findings must be assembled into report format using scales or degrees of likelihood.

Did the suspect write it?	Scale	Conclusion
YES	9	Identification
	8	Highly probable - did write
	7	Probable — did write
	6	Indications — did write
INCONCLUSIVE	5	No conclusion
NO	4	Indications — did not write
	3	Probable — did not write
	2	Highly probable – did not write
	1	Elimination

**Table 3.12: “Conclusions on resemblance between questioned and known writings”, from:  
McMenamin (2002, Section 6.4.5)**

Table 3.12 above depicts a report where the suspect is confirmed or ruled out as author. It gives a scale from 1 to 9, 1 being innocent and 9 being confirmed as the likely offender.

Did one author write it?	Scale	Conclusion
<b>YES</b>	9	Definite — one writer
	8	Highly probable — one writer
	7	Probable — one writer
	6	Indications — one writer
<b>INCONCLUSIVE</b>	5	No conclusion
<b>NO</b>	4	Indications — more than one writer
	3	Probable — more than one writer
	2	Highly probable — more than one writer
	1	Definite — more than one writer

**Table 3.13: “Conclusions on consistency within known and questioned writings”, from:  
McMenamin (2002, Section 6.4.5)**

Table 3.13 depicts a similar system as the previous example, in that it uses a scale to depict another regarding the probability of the author in question as the potential offender who wrote the questioned material.

### **3.7.5 Kings College London Approach to Text Analysis**

Kings College London has created a type of methodology for text analysis where little is known about the material in question.

#### **i. Types of Text Analysis**

According to Kings College (2007), text analysis can fall into one of three categories. These are as follows:

- **Concording** – This is the primary method of analysis for this type of work
- **Content analysis** – This is often utilised when performing qualitative analysis, and its use involves extracting words and terms from the text in question for concording or statistical analysis
- **Statistical analysis** – Here frequency lists and distributions are created by applying mathematical transformation to the text in question



## **ii. Factors to take into account when performing Concordance**

According to Kings College (2007), the more the investigator knows about the text in question, the more efficient the analysis will be. They add that the focus when performing concordance is the technique, which will allow the investigator to acquire data quickly. The following points should be considered when investigating a text.

- Genre - This describes the way in which a person speaks and writes e.g. legal or poetic. Questions the investigator should ask himself are whether the documents were originally written or taken down some other way, what features can be expected to be present in this type of text and can they be spotted?
- Rhetoric and vocabulary – The genre shapes the way in which the vocabulary is used. This being said, the investigator can look for the repeated usage of certain lexical categories.
- Social or psychological circumstances – The circumstances around which the text was written could possibly be relevant to the investigator as it could depict a particular writing style.
- Historical circumstances – Historical characteristics of the text in question are of paramount importance. A writer's historically-defined vocabulary characteristics will allow the investigator faster location of key terms.
- Nature of the artefact – Should text be taken from another source, it may be beneficial to understand the type of source used as it will aid in providing valuable historical data for the analysis process.

## **iii. The Analysis Process**

Kings College (2007) states that it is important for the investigator to approach the analysis process with an open mind as this will allow for the acquisition of important data that was least expected. The analysis procedure follows three steps which can be used individually or in combination, as described below:

- High Frequency Words – Going through a list of the most frequently used words can be useful when trying to identify words that are characteristic of the text in question. Frequency analysis has many other uses for corpus analysis.

- Collocations – This is the method of determining what words tend to be found together or within a certain proximity to another word. It is understood that repeated collocations are more reliable indicators of meaning than single word repetitions. A program that can be used to perform this type of analysis is known as Monoconc and is described in section 3.6.2.1.1.
- Concordance Analysis – This system involves giving understanding of the immediate environment of a specified word (X number of words before and after the specified word). By using this system, a writer's unique style can be identified.

### **3.8 Summary**

This chapter began with providing an explanation of the types of linguistic analysis methods in existence. It then proceeded to explain what linguistic variation is about as well as the usage of language markers and the idiolect. In order to demonstrate the effectiveness of forensic linguistic analysis, relationships between DNA and language stylistics were assessed, proving that language usage is unique among different people and can be used to identify a particular person. Lastly, information was provided regarding how a framework for authorship identification analysis should be derived taking into account various pertinent factors.

The following chapter explores the creation of the conceptual framework for ethical academic writing as well as explains the realm of cyber forensic auditing. Finally, the chapter will include a hypothetical scenario involving a case of alleged plagiarism at a university, to demonstrate the effectiveness of the developed framework.

## *Chapter Four*

### A CONCEPTUAL FRAMEWORK FOR ETHICAL ACADEMIC WRITING

#### **4.1 Introduction**

This chapter aims to show how the conceptual framework designed in figure 4.2 can be applied to a breach of ethical academic writing. It is important to note that the framework is a generic template that can be applied to many types of cyber forensic situations. The legal aspects mentioned in section 2.15.4 of the FORZA framework can be included in this conceptual framework; however, due to the limitations of this research, these factors will be addressed in a later study.

The chapter will begin with a general understanding of where the elements of cyber forensics and linguistic forensics are in relation to each other, as well as in relation to their parent, the realm “Forensic Auditing”. Once this is explained, the conceptual framework will be presented, along with a detailed explanation of how it works.

On completion of this, a “mock” scenario regarding a breach of ethical academic writing at an academic institution will be assessed. This will be done to show how the conceptual framework can be applied in a structured manner to more efficiently deal with a detected breach of ethics.

## 4.2 Cyber Forensic Linguistics

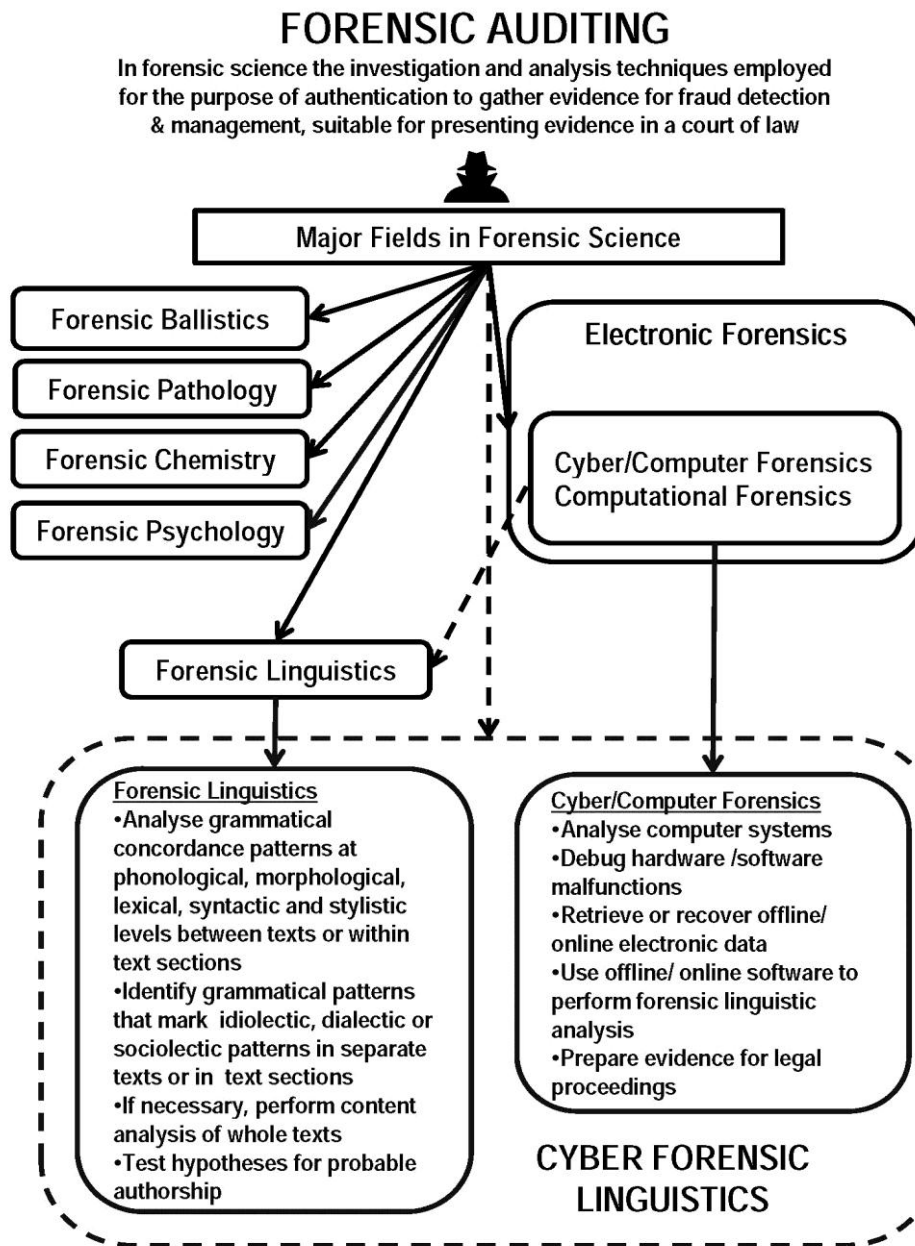


Figure 4.1: Cyber Forensic Linguistics as a Component in the Greater Realm of Forensics

Figure 4.1 above depicts the realm of forensic auditing with a specific focus on cyber forensics as a sub-component. The diagram shows six of the many different types of forensics that exist in the greater realm of forensic science: forensic ballistics, pathology, chemistry, psychology, linguistics and electronic forensics. The element “electronic forensics” contains a sub-element known as cyber/computer forensics, which shares certain characteristics with forensic linguistics. These similar traits are: utilising computers to perform document and/or word analysis e.g.

concordance analysis, function word analysis and quantitative analysis. The dotted line that links forensic linguistics to computer forensics depicts this relationship and represents an indirect link between the two elements. This new relationship creates a new realm of linguistics and forensics known as “cyber forensic linguistics”. The element “cyber forensic linguistics” is depicted above and shows the various functions of forensic linguistics and those of cyber/computer forensics comparatively.

### **4.3 General Conceptual Framework for Forensic Linguistic Analysis**

Figure 4.2 on the following page depicts the complete forensic linguistic investigation lifecycle. This framework follows a set and structured sequence of steps to accomplish the investigation process. The core steps are identified below and discussed in greater detail after figure 4.2 has been presented.

The first step is the initial contact between the client and investigator; here the client comes across an irregularity and hires the forensic investigator to perform the investigation (Initiating Step 1). Once the client has contacted the investigator, the investigator draws up the mandate letter that most importantly indicates the fee for the investigation, the boundaries of the investigation and client requirements for the investigation, amongst other things (Initiating Step 2). Once this letter has been drafted, it is given to the client to sign. At this stage the investigator goes through the requirements in the mandate letter with the client so as to confirm their understanding before the client signs the letter (Initiating Step 3).

Once the mandate letter is signed, the investigation can proceed according to the core steps (Initiating Step 5). The core steps consist of five processes: Identify the Problem (Step 1), Gather Data (Step 2), Analyse the Data (Step 3), Produce Results (Step 4) and lastly, Create Report (Step 5). When performing the core steps, it is important to take into account the legal aspects presented in section 2.15.4 of the FORZA Framework.

The FORZA framework aims to bridge the gap between the technical and legal aspects of a forensic investigation, as many legal experts consider technical procedures difficult to learn. The framework allows for the incorporation of business, system and legal aspects, and gives investigators the ability to view the underlying concepts (confidentiality, integrity and availability) of the core IT security fundamentals with relative ease. An investigation process comprises various steps

and phases; this framework aims to allow for certain roles and functions to be handled by the same person and weaves in a more proactive role for the legal adviser and prosecutor. Finally, by allowing the legal advisor to play a more active role in the investigative process, the probability of the case being dismissed on legal technicalities is greatly reduced.

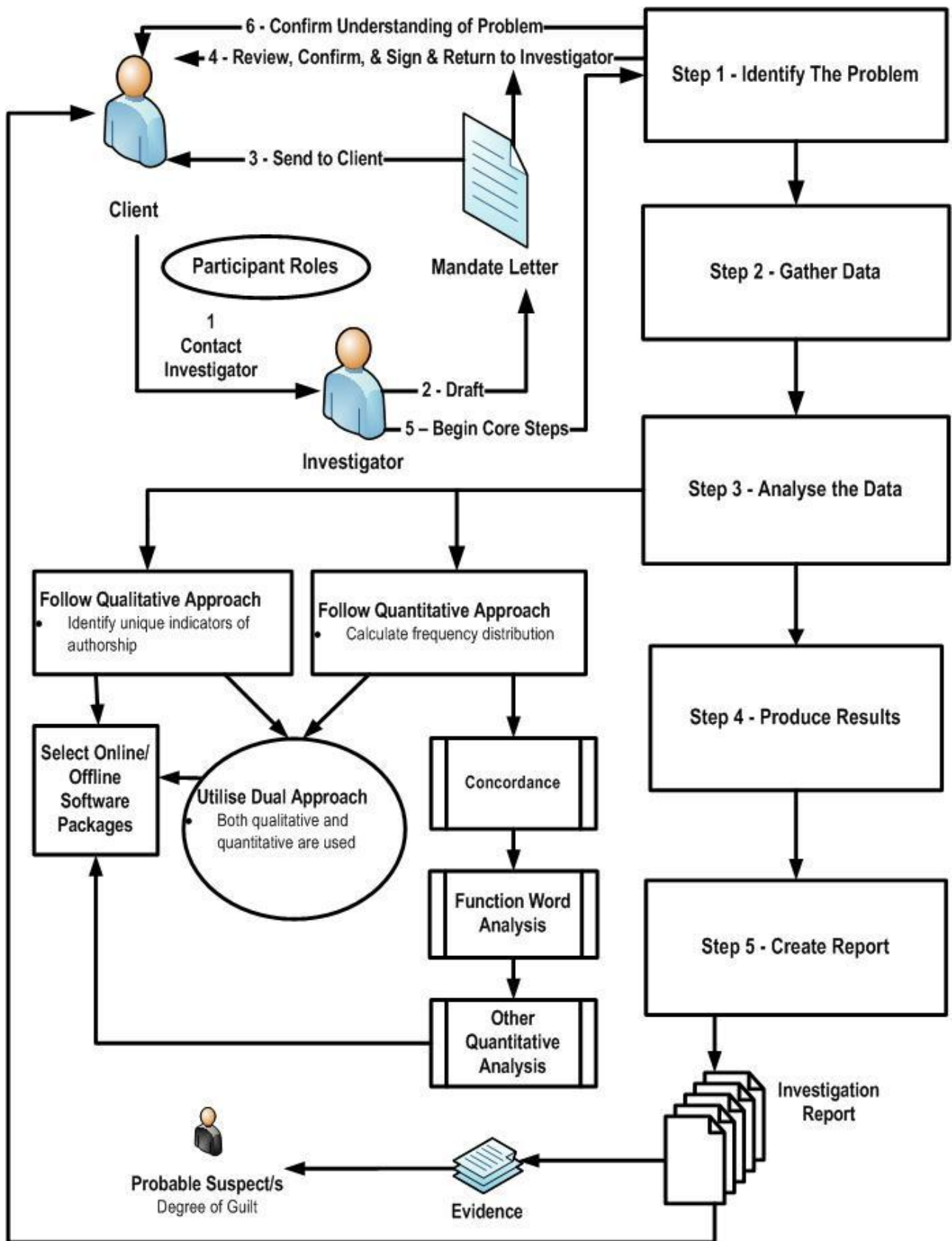


Figure 4.2: General Conceptual Framework for Forensic Linguistic Analysis

**i. Core Step 1 – Identify the Problem**

Here the creation of the mandate letter and gaining agreement and consensus on the requirements are the focus. The investigator reviews the case and discusses her understanding of it with the client. The investigation research questions are created in this phase as well.

**ii. Core Step 2 – Gather the Data**

This phase involves the collection of all relevant data pertaining to the case. Evidence must be “bagged and tagged” so as to preserve the chain of custody and to ensure the evidence does not get contaminated, thereby rendering it invalid for use in a court of law. Original evidence must be duplicated and the copies used to work from. For digital evidence, the suspect’s hard drive can be “imaged” using specialised software packages that ensure CRC and MD5 values correspond, e.g. Encase. In the case of printed and handwritten material, these can simply be photocopied. It is important to preserve the original evidence and to ensure it is not changed at all. When performing concordance, function word and other quantitative analysis techniques, texts that are in hardcopy (printed) format can be scanned and run through OCR processes to convert them to electronic format. This has major benefits in that the electronic data can then be analysed using many of the available software tools. This will convert a potentially laborious process into a much shorter and less mistake-prone system.

**iii. Core Step 3 – Analyse the Data**

Software programs to be used for the analysis process must be analysed and chosen. The packages that are chosen will depend on the type of data available, time constraints, budget as well as the client’s requirements. Concordance tools, function word analysis tools and quantitative methods can be employed for the analysis. A variety of software tools exist for each of these types, both online and offline.

The data can be analysed by following a qualitative or quantitative approach. However, for best results, a mixed/dual approach is recommended. This dual approach is simply the combination of both analysis methods in order to produce more accurate results. When going about the analysis, a variety of software programs exist to facilitate easier and quicker analysis of the data. Since requirements and technology are constantly changing, these packages will need to be analysed and re-analysed whenever an investigation of this nature presents itself, in order to



determine which are most suitable for the data and case in question. It is also important to take into account the client's requirements when deciding on the software packages to use, as paying for a fully-featured package when realistically only one or two aspects will be used, is not ethical or feasible.

**iv. Core Step 4 – Produce Results**

The results of the variation must be compared to both the known and questioned texts in order to determine the degrees of similarity or deviation. This will show how alike two pieces of text are; conclusions can then be drawn based on degrees of similarity (see tables 3.12 and 3.13).

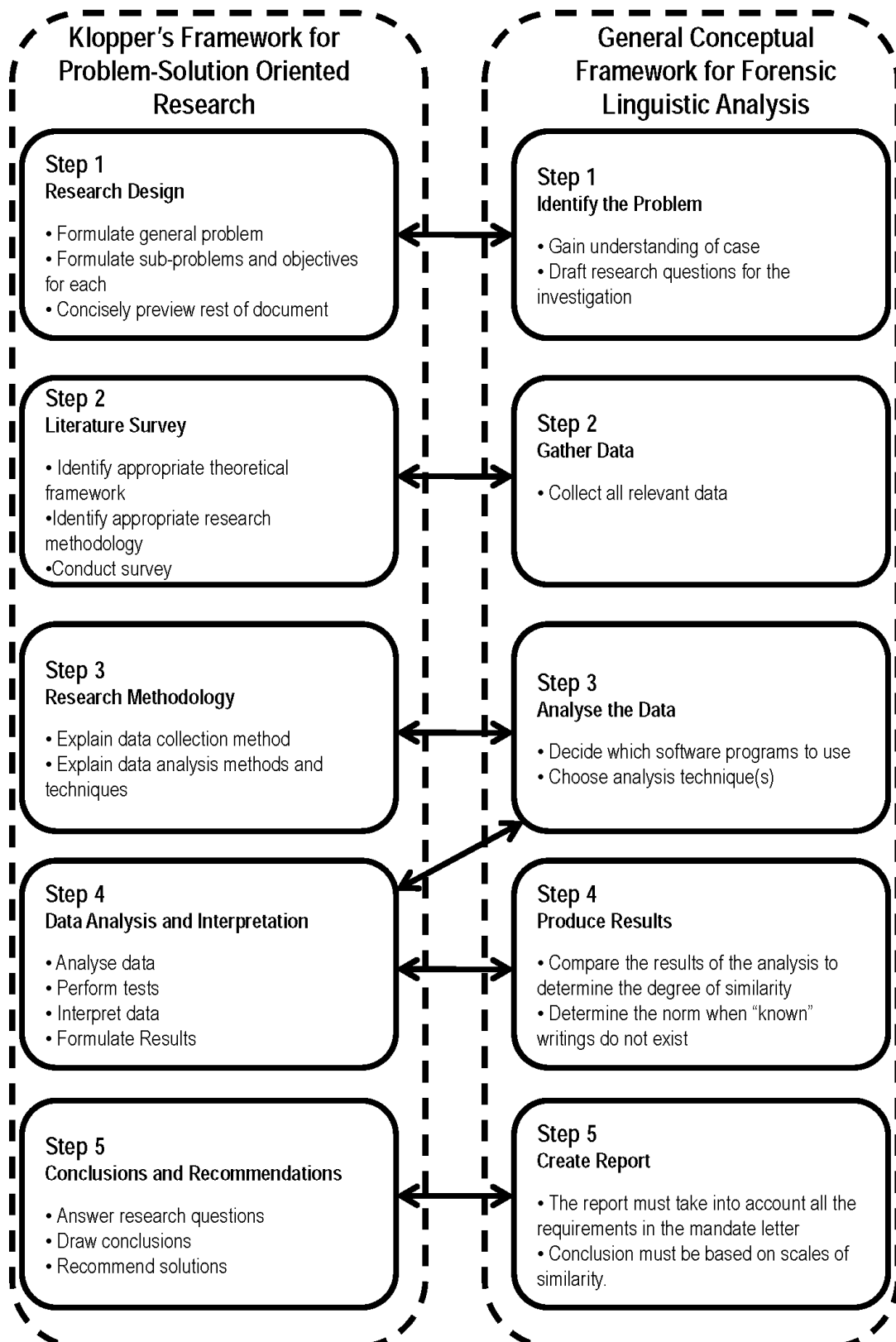
**v. Core Step 5 – Create Report**

The report must be created taking into account all the requirements of the mandate letter. The conclusions in the report must be based on scales of similarity and must indicate the degree of guilt of the suspect/s. Any acquired evidence must also be presented with the report.

This framework allows for a structured and more efficient response when going about a cyber forensic investigation. It is important to note that this framework is a conceptual model and has not been practically tested due to the limitations of this study.

**4.4 Similarities between the General Framework for Problem-Solution Oriented Research and the General Conceptual Framework for Forensic Linguistic Analysis**

The similarity between problem-solution oriented research (figure 1.1) and forensic linguistic analysis (figure 4.2), will no doubt not have been lost on the reader. These similarities are summarised in figure 4.3 below, and thereafter briefly discussed.



**Figure 4.3: Similarities between Klopfer's framework for problem-solution oriented research and the general conceptual framework for forensic linguistic analysis**

Firstly, it is important to note that Klopper's framework for problem-solution oriented research can be characterised as a generic process. It is a tried and tested system for performing problem-solution based research (research aimed at solving a type of issue) and forms the basis of the technique used in this study.

Secondly, it is important to note that Klopper's framework for problem-solution oriented research as well as the proposed general conceptual framework for forensic linguistic analysis, are empirical by nature in view of the fact that data are gathered, then subjected to procedures of systematic analysis in order to obtain results on which objective, defensible conclusions can be based.

Finally, figure 4.3 depicts the shared aspects between the two models. It is clear that there are many similarities and that the main differences can be attributed to the fact that Klopper's framework is a more generic process whereas the general conceptual framework is more specific, in that it pertains solely to the discipline of cyber forensics. The similarities are presented to show that Klopper's framework is a system for solving research-based problems, whereas the general conceptual framework exists for solving forensic-based problems. The similarities also prove that the newly developed framework is a generic system for forensic investigatory problems. It is anticipated that the principles of acoustic phonetics (section 3.2.2) will assume increased importance when forensic auditing has to incorporate speech to text analysis as part of the process for establishing authorship.

#### **4.5 The Framework Applied to an Hypothetical Scenario**

This section will provide a hypothetical scenario involving an investigation into plagiarism to explain how the framework developed in section 4.3 could be applied to detect a breach of ethics at an academic institution. Please refer to figure 4.2 and the subsequent step-by-step explanation of the procedure.

##### **4.5.1 Initiating Step 1**

University X contacts investigator regarding a potential case of plagiarism at the university. The university feels the incident was plagiarism, as the student's assignment is very similar to that of an article on the Internet.

#### **4.5.2 Initiating Step 2**

The investigator discusses the terms of the investigation with the university and draws up the mandate letter detailing all the requirements of the investigation, the fee structure, transferral of authority to run interviews and interrogations, time frames and other aspects that he/she may deem important.

#### **4.5.3 Initiating Step 3**

The investigator sends the mandate letter (containing various terms and conditions) to the university and discusses his/her understanding of the case with the relevant university authorities.

#### **4.5.4 Initiating Step 4**

The university confirms and signs the mandate letter and returns it to the investigator, thereby agreeing to the investigator's terms and conditions.

#### **4.5.5 Initiating Step 5**

The investigator can now commence with the core investigation steps.

#### **4.5.6 Core Step 1 - Identify the problem**

In this step the investigator aims to gain an understanding of the issues at hand and draws up research questions to achieve certain milestones in the investigation of the case. The investigator will also identify all pertinent role players with regard to this case, which in this instance would be the head of school of the student's faculty, the head of the university's internal investigations unit and finally, the author of the work that the university suspects the student of plagiarising from. The relevant research questions will be: has the student been accused/convicted of plagiarising before? What is the student's current academic situation (i.e. is he/she near exclusion, therefore potentially becoming desperate)? Is this course critical for the student to pass? Can we obtain another assignment belonging to the same student (function word and content analysis)? Is the student aware of the university's rules regarding plagiarism? What are the university's rules regarding plagiarism?

#### **4.5.7 Core Step 2 – Gather Data**

Here all relevant data must be collected. Evidence must be collected in a manner legally acceptable so as not to contaminate the evidence or damage the chain of custody. Original evidence must be duplicated where possible and copies used to work from. Typed texts can be run through OCR processes to make them editable and compatible with the various corpus based analysis software in existence.

In the case of the university, the investigator will request the student's assignment in question from the head of school as well as the website article that the student supposedly plagiarised from. The investigator can also request another assignment belonging to the student if possible to use for comparative purposes. Other data the investigator could also request include the student's academic record; interviews could also be scheduled with his/her lecturers and mentors. If the assignments are in hard copy printed format, the investigator will run them through OCR processes to transform them into editable electronic format. All hard copy evidence must be photocopied and the copies used to work from. The interviews can be recorded using recording devices and categorised. All evidence must be kept in evidence bags in a secure safe/vault.

#### **4.5.8 Core Step 3 – Analyse the Data**

This step involves the analysis of the data acquired in step 2. The case of alleged plagiarism will require approaches dependent on the type of data available. In this scenario the investigator has a hard copy of the assignment in question as well other material the student has handed in before the incident. Suitable tools will have to be decided on, and in this case *Turnitin.com* can be used. The assignment in question can be run through an OCR process to get it into editable electronic format. In this format the investigator can then feed it into *Turnitin.com*. Another form of analysis that can be performed is function word analysis, as the investigator has another piece of work that was done by the student in the past. Word frequencies can be derived from the assignment in question, the student's other piece of work as well as the web site that the student is accused of plagiarising from.

#### **4.5.9 Core Step 4 – Produce Results**

In this step the results of the analysis done in the previous step will be analysed and compared. The results of the word frequency calculations can be compared to each other. If it is found that the frequency of the assignment in question is very close to the frequency of the web site, then one can deduce that the student did in fact plagiarise from the site. However, if the frequencies differ and the frequency of the student's previous assignment differs vastly from the frequency of the assignment in question, then one can conclude that the student is innocent. The other option of using *Turnitin.com* can also be looked at, as it will output colour coded results should it find any plagiarism.

#### **4.5.10 Core Step 5 – Create Report**

Once the results have been analysed, the investigation report can be drawn up. This report will contain scales or degrees of similarity regarding the student's assignment in question and its likeliness to the web site. Other scales can be drawn up regarding comparisons to factors such as *Turnitin.com* and the student's other assignment. The report must contain a list of terms that are commonly used as well as their definitions and explanations. Finally, the report must cater for all the university's requirements as per the agreement in the mandate letter. Any evidence must be indexed and presented with the report in an easy to understand manner.

### **4.6 Summary**

This chapter began with an explanation of where cyber forensics and linguistic forensics are in relation to each other and their parent, forensic auditing. Diagram 4.1 depicts this relationship and additionally reveals the indirect relationship between various elements that have led to the creation of the new discipline of linguistics known as "cyber forensic linguistics"

The chapter then proceeded with the creation of the general conceptual framework for forensic linguistic analysis. This framework is the culmination of various aspects of this research and was demonstrated in this chapter through a hypothetical scenario involving a case of alleged plagiarism at a university (section 4.5).

In Chapter 5, the Conclusions and Recommendations, the writer will review the research questions and provide an explanation of the extent to which these questions were answered, as well as details regarding the areas in which the research can be furthered.

## *Chapter Five*

### CONCLUSIONS AND RECOMMENDATIONS

#### **5.1 Introduction**

This chapter contains a concise overview of the previous chapters as well as a detailed description of the extent to which the research questions detailed in Chapter 1 were answered. The chapter will end with the conclusions and recommendations emanating from the study as a whole.

#### **5.2 Recap/overview of the Research**

The study assumes the reader to have knowledge of the subject matter equivalent of that of a lay-person. This being said, the study begins each chapter by explaining the basics that form part of much more complex systems later on. The following sections give a general review of the chapters in this study.

##### **5.2.1 Chapter 1 - Introduction**

This chapter provided a general introduction to the study as well as a detailed motivation for why the research is being done. This section also breaks down the major problem statement into subproblems, formulates an objective for each subproblem and addresses each on their own.

##### **5.2.2 Chapter 2 – Literature Review**

In this chapter literature on the various aspects of forensics, electronic computing, communication devices and plagiarism, were reviewed. These topics included aspects such as frameworks and how to apply them, and legal aspects that should be taken into account during forensic auditing. The issue of plagiarism, which is central to this study, was addressed in detail in this chapter. Plagiarism was further analysed by exposing the various types of plagiarism in existence as well as how to identify them. Various remedies and techniques to combat the problem of plagiarism were included in this chapter as well.



The usefulness of using syntactic patterns to establish authorship was also discussed, along with details regarding the various types of lexical categories and how to use word frequency analysis to establish authorship. Authentication methods were explained, in particular the MD5 and CRC values. An explanation of the cyber forensic investigation process was included to give an understanding of some of the factors that were taken into account in the creation of the conceptual framework. Electronic tools that have been specifically created to combat plagiarism were presented. The concept of “ethics” and how it impacts on a person’s perception on plagiarism was provided in this chapter.

### **5.2.3 Chapter 3 – Forensic Linguistics**

Chapter 3 dealt mainly with the linguistic aspects of the research. The stylistic analysis from Chapter 2 forms the basis upon which these concepts were built. The chapter begins with linguistic analysis, variation and the various analysis techniques in existence. These items provide the much needed theoretical background for the systems included in the conceptual framework in Chapter 4.

The unseen relationships between forensic linguistics and various other aspects of forensic auditing were shown here. This section provided the basis for an understanding that many important aspects of forensic auditing that we presently assume to be unrelated, are in fact related in some way, allowing the researcher to draw upon new concepts to further enhance analysis. The concept of a “norm” was examined, as well as how it is derived. Theory justifying the processes and aspects included in the conceptual framework as well as the mandate letter and evidence handling, were included in this chapter.

### **5.2.4 Chapter 4 – A Conceptual Framework for Forensic Linguistic Analysis**

In this chapter, the conceptual framework for forensic linguistic analysis was created. Information regarding the new discipline of cyber forensic linguistics was also drawn up. The chapter concluded with a hypothetical scenario involving a case of alleged plagiarism at a university. This was done to demonstrate the usage and effectiveness of the designed conceptual framework.

## **5.3 Research Questions Revisited**

### **5.3.1 Introduction**

This section explains the extent to which the research questions presented in Chapter 1 were answered. Taking as point of departure, the fact that objectives are derived from problem statements, and that research questions are in turn derived from objectives, it should be clear that the extent to which one manages to answer a particular research question is the extent to which one has met its objective, which is the extent to which one has managed to solve the problem initially identified. Therefore, in the rest of this section, the researcher will indicate to what extent he has been able to answer the three specific questions raised as the initial research problems.

The following are the research questions from Chapter 1. It is important to note that this study follows a problem-solution oriented approach as described in the previous paragraph.

### **5.3.2 General Research Question**

Could intellectual property theft constitute a breach of ethical writing?

Answer – To concisely describe ethical writing is to conclude that it embodies the most important factors that describe the manner in which text should be composed. In Section 2.4 the concept of honesty is of greatest importance when dealing with ethical writing, whereas plagiarism can be described as the worst violation of it. Plagiarism is the theft of another's intellectual property and comes in many different types and forms. These can range from deliberate, all the way to the accidental theft of another's work as detailed in section 2.7. It is clear that a breach of intellectual property is a violation of ethics and therefore also a breach of ethical writing.

### **5.3.3 Subquestion 1**

What roles could various forms of text analysis play as forensic tools in determining the quality of ethical academic writing?

Answer – There are many tools and techniques in existence that can be used to determine the quality of ethical academic writing. These tools are oriented around authorship identification as well plagiarism detection. The first set of tools that can be looked at are the anti-plagiarism tools; there are many of these in existence, some of which are described in section 2.6.7 These tools can fall into two different categories, either online or downloadable software tools.

The second set of methods are more specifically centred around corpus analysis and are comprised of either concordance, function word analysis or other quantitative analysis. An important feature of corpus analysis is that of annotation, whereby tags are added to the textual items within the corpus in order to identify linguistic patterns. These tags are based on categories of text annotation described in further detail in section 2.7.2. Modern content analysis techniques such as function word analysis allow for the creation of word frequency distributions (section 2.17) and form part of corpus analysis. Using software tools is a definite benefit over performing manual analysis, with speed and accuracy of processing data forming only part of the reason why one should always attempt to use one of the software tools available. A further motivation is the fact that processing power of modern computers is always on the increase. We can assume that with this increase in processing power, a decrease in the time taken to perform corpus analysis using these electronic tools will be evident. Corpus analysis tools exist for almost all types of text analysis and these tools are constantly being updated and improved.

Sections 2.2 to 2.4 dealt with the complexities of assigning authorship, showing that one cannot depend on non-semantic and non-linguistic arguments to determine authorship. Section 2.3 specifically highlights the role language analysis plays in forensic auditing: the reader is introduced to the process of breaking down sentences into their core constituents as well as the concept of word frequency analysis, both of which are vital to corpus analysis, one of the underlying concepts of this study. Corpus analysis methods can be performed manually; however, it is not feasible to perform these analysis techniques on large corpora as it is both very time-consuming and expensive. Sections 3.4 and 3.6 describe the various components of corpus analysis, providing samples and features of the various tools that are in existence to perform these types of analysis.

### **5.3.4 Subquestion 2**

What would be the elements of a conceptual framework for cyber forensic auditing that accommodates the study of ethical writing as part of cyber forensics?

Answer – In order to provide an accurate and complete answer to this problem, many factors needed to be taken into consideration. Firstly, existing frameworks needed to be assessed in order to determine the functions and sequences of operations. Aspects of the FORZA framework (section 2.15) were chosen for this purpose as they dealt specifically with legal aspects that most other frameworks neglect to take into account. Other important aspects were drawn from section 2.11.1 (processes and factors involved in a digital forensic investigation), as well as section 2.13 (digital crime scene investigation framework). These frameworks were assessed and used in the derivation of the conceptual framework in figure 4.2.

Secondly, the analysis aspects needed to be determined. These were specifically the corpus analysis (section 2.8.1) tools and methods that could be used within a framework for cyber forensic linguistic analysis. The techniques employed were content analysis (section 2.17), concordance analysis (section 3.6.2.1.2), function word analysis (section 2.8.3) and various other tools (section 3.6.2).

### **5.4 The Importance of Forensic Linguistics in Cyber Forensic Analysis**

Forensic linguistics is an aspect the forensic investigator cannot afford to underestimate when going about a case of authorship identification. This research shows that forensic linguistic analysis is a highly effective and reliable system. As technology progresses and advances, so does the potential to misuse it. The discipline of cyber forensics is ever growing due to the progression of technology. While linguistic analysis has been in existence well before computers were around, in the modern era this field has been extended to cyber forensics as there is an abundance of software tools available to perform linguistic analysis.

## **5.5 Conclusions and Recommendations**

### **5.5.1 Conclusions**

By demonstrating that the study enabled the researcher to answer the two research questions that were posed about the problem under investigation, it can be concluded that solutions have been found for the two major problems regarding plagiarism, namely (1) the roles various forms of text analysis could play as forensic tools in determining the quality of ethical academic writing, and (2) the elements that should comprise a conceptual framework for cyber forensic auditing that accommodate the study of ethical writing as part of cyber forensics. These two research problems go hand-in-hand as the various text analysis tools can now be implemented as part of an element of the conceptual framework developed in this study. The importance of this conceptual framework cannot be understated as it provides all the steps and functions required to effectively respond to a case of suspected plagiarism, and its creation entailed the study of several different frameworks and forensic investigatory methods.

Besides the elements of cyber linguistics that were identified, a number of offline and online tools have been provided for determining authorship in source texts that have been suspected of being plagiarized, as well as for identifying specific sections in the plagiarised text that have been derived from one or more source texts. It was also shown that just as DNA can be used to prove individual uniqueness, so can an individual's language style or idiolect. The linguistic methods that can be utilised during a forensic audit of this nature have also been identified, namely (1) word frequency analysis, (2) concordance and (3) other quantitative analysis methods. It is important to note that no single method alone can be regarded as effective enough to prove authorship. These techniques must be used together to ensure an honest outcome that can be regarded as reliable in a court of law.

The combination of the indirectly related aspects of both cyber forensics and electronic forensics has led to the creation of a new discipline in the realm of forensic auditing. This discipline, known as "cyber forensic linguistics", has the specific purpose of combining the electronic tools and techniques available and integrating them with the various corpus analysis methods for authorship identification. This

process results in computerised corpus analysis methods that allow for faster and more accurate analysis of linguistic data.

The study described in detail every step that the conceptual framework comprises, as well as the possible role players. It is important to note that the conceptual framework is not limited in usefulness to cases of alleged plagiarism, but can also be applied to various other forms of computer misuse as well as authorship identification cases.

### **5.5.2 Recommendations**

A conceptual framework is a theoretical tool that has not been practically tested. In subsequent research, the usefulness of this framework in a practical situation must be determined. There is still much room for enhancement of the developed conceptual framework. These various steps in the framework should be further refined through practical testing and a deeper analysis of the various role players in each step performed. The various software packages can be acquired and individually tested to determine the practical effectiveness of each, as well as the analysis of new packages that are in existence.

Legal aspects highlighted in the FORZA framework should be added to the conceptual framework and tailored specifically for the country/area in use. Detailed research into the legal requirements regarding a forensic linguistic investigation can be undertaken, and similar cases that have been tried around the world can be compared and analysed.

Finally, the forensic framework should be expanded and tailored for specific situations, for example if the investigation surrounds cellular telephone SMS (Short Message Service) technology. Since this study dealt mostly with syntax and lexicon, the other branches of linguistics, namely, phonetics, phonology and morphology, could be further assessed to determine effectiveness as being part of the forensic framework designed in this study. Specific linguistic traits tailored for a South African perspective could be created. This would be a form of a corpus containing words and typical word pattern usage that are specifically found in South Africa. In the final analysis, with a word set such as this, the analysis phase of the framework

would proceed at a much quicker and efficient pace for certain cases of authorship identification.

## BIBLIOGRAPHY

- Argamon, S and S, Levitan (2005). Measuring the Usefulness of Function Words for Authorship Attribution. *Proceedings of ACH/ALLC Conference 2005*, Victoria, BC, Canada, June 2005. Available at <http://lingcog.iit.edu>. [Accessed on 14 August 2007 10:30 AM].
- Article Checker (2008). Article Checker. Available at <http://www.articlechecker.com/>. [Accessed on 28 October 2009 13:53].
- AskOxford.com (2007a). Using The Oxford Corpus. Available at <http://www.askoxford.com/>. [Accessed on 27 August 2007 13:20].
- AskOxford.com (2007b). *Oxford Dictionaries*. Available at <http://www.askoxford.com/results/?view=dict&freesearch=stylistics&branch=13842570&textsearchtype=exact>. [Accessed on 21 August 2007 14:20].
- AskOxford.com (2007c). *Compact Oxford English Dictionary*. Available at [http://www.askoxford.com/concise\\_oed/linguistics?view=uk](http://www.askoxford.com/concise_oed/linguistics?view=uk). [Accessed on 21 August 2007 15:12].
- AskOxford.com (2008). *Compact Oxford English Dictionary*. Available at [http://www.askoxford.com/concise\\_oed/intellectualproperty?view=uk](http://www.askoxford.com/concise_oed/intellectualproperty?view=uk). [Accessed on 14 August 2008 12:51 AM].
- Athelstan (2007). *Concordances and Corpora*. *Athelstan Publishers*. Available at <http://www.athel.com/mono.html>. [Accessed on 27 August 2009 13:12 PM].
- Barnbaum, C (2002). *Plagiarism: A Student's Guide to Recognizing It and Avoiding It*. Available at [http://www.valdosta.edu/~cbarnbau/personal/teaching\\_MISC/plagiarism.htm](http://www.valdosta.edu/~cbarnbau/personal/teaching_MISC/plagiarism.htm) [Accessed on 31 May 2005 1:27 PM].
- British Medical Journal (BMJ) (2006). Detecting plagiarism: Google could be the way forward. Available at <http://www.bmj.com/cgi/content/full/333/7570/706-b> [Accessed on 14 August 2007 12:51 AM].



- Branford, W (ed). (1991). *The South African Pocket Oxford Dictionary*. Cape Town Oxford University Press.
- Broucek, V and Paul Turner. (2004). Computer Incident Investigations: e-forensic Insights on Evidence Acquisition. In U.E. Gattiker (Ed.), *EICAR 2004 Conference CD-rom: Best Paper Proceedings*. Available at <http://forensics.utas.edu.au/files/EICAR2004.pdf> [Accessed on 9 July 2005 23:30].
- Bunting, S and Wei, W (2006,p157). *EnCase Certified Examiner Study Guide* ISBN – 0-7821-4435-7.
- Cambridge (2007). *Dictionary*. Available at <http://dictionary.cambridge.org/> [Accessed on 27 August 2007 13:11].
- Canexus (No Date). Eve2. Available at <http://www.canexus.com/support.shtml> [Accessed on 5 October 2009 21:14].
- Carrier, B and Spafford, E (2004). An Event-Based Digital Forensic Investigation Framework. *Presented at the Digital Forensic Research Workshop (DFRWS)*.
- Concordance (2009). Available at <http://www.concordancesoftware.co.uk/>. [Accessed on 27 August 2009 19:32].
- CopyCatchGold (2009). Available at <http://www.copycatchgold.com/>. [Accessed on 3 October 2009 09:21 AM].
- CopyScape (2009). Available at <http://www.copyscape.com/>. [Accessed on 28 October 2009 20:42].
- Coxhead, P (2007). A Referencing Style Guide – *University of Birmingham*. Available at <http://www.cs.bham.ac.uk/~pxc/refs/index.html>. [Accessed on 21 April 2007 10:44 AM].
- Davidson, Robert, Malcom Monro and Detmar Straub (2003). AIS Code of Research Conduct. *Association For Information Systems*. Available at <http://home.aisnet.org>. [Accessed on 13 October 2008 12:40].
- Dupli Checker (2009). Available at <http://www.duplichecker.com/>. [Accessed on 27 October 2009 18:57].

- Dutton, D (1992). *Umberto Eco on Interpretation - Johns Hopkins University Press*. Available at [http://www.denisdutton.com/eco\\_review.htm](http://www.denisdutton.com/eco_review.htm). [Accessed on 10 April 2010 13:20].
- Eco, U (1990:1). *The Role of the Reader*. ISBN – 0 09 146391 2
- Eco, U (1994:2). *The Limits of Interpretation*. ISBN – 0 253 31852 1
- Glatt Plagiarism Screening Program (No Date). Available at <http://www.plagiarism.com/screen.id.htm>. [Accessed on 16 October 2009 18:21].
- Hexham, I (1999). *The plague of plagiarism - Department of Religious Studies The University of Calgary*. Available at <http://c.faculty.umkc.edu/cowande/plague.htm>. [Accessed on 25 August 2008 10:59 AM].
- Holcombe, J (2007) – *Linguistics: A brief history*. Available at <http://www.textetc.com/theory/linguistics.html> [Accessed on 21 October 2009 10:05 AM].
- Ieong, R (2006). *FORZA – Digital Forensics Investigation Framework that Incorporates Legal Issues*. caps Available at <http://dfrws.org/2006/proceedings/4-Ieong.pdf>. [Accessed on 27 August 11:53 AM].
- Jplag (No date). *University of Karlsruhe*. Available at <https://www.ipd.uni-karlsruhe.de/jplag/>. [Accessed on 12 October 2009].
- ISACA (2004). *IS Auditing Guideline - Computer Forensics. Document G28* Available at [www.Isaca.org](http://www.Isaca.org)
- Kilgarriff, A (1997). *Using Word Frequency Lists to Measure Corpus Homogeneity and Similarity between Corpora*. caps Available at <http://www.itri.brighton.ac.uk>. (ITRI-97-07).
- Kings College of London (2007). *Fundamentals of the digital humanities method in text-analysis: An introduction*. Available at <http://www.cch.kcl.ac.uk/legacy/teaching/av1000/textanalysis/method.html>.

- [Accessed on 11 August 2009 21:55].
- Klopper, R (2008). Problem-Solution Oriented Research. Unpublished document.
- Knight, E and Jason, M and (2009). Safeassign. Available at <http://wiki.safeassign.com>. [Accessed on 1 October 2009 14:54].
- Leech, G (2004). Developing Linguistic Corpora: a Guide to Good Practice. Adding Linguistic Annotation. *Lancaster University*. Available at <http://ahds.ac.uk/guides/linguistic-corpora/chapter2.htm>. [Accessed on 28 August 2007 13:58].
- Marcella, A and Robert, G (2006). Cyber Forensics. Available at <http://www.cyber-forensic-analysis.com/>. [Accessed on 15 September 2008 12:31 AM].
- McEnery, T and Andrew, W (No Date). Corpus Linguistics. Available at <http://bowlandfiles.lancs.ac.uk/monkey/ihe/linguistics/corpus2/2fra1.htm>. [Accessed on 28 August 2007 12:40 AM].
- McMenamin, G (2002). Forensic Linguistics : Advances in Forensic Stylistics  
*ISBN 0-8493-0966-2*
- Missikova, G (2003) – Linguistic Stylistics Available at <http://www.ff.ukf.sk/kaaa/staff/missikova/Linguistic%20Stylistics%20part%201.pdf>  
[Accessed on 14 August 2007 9:43 AM].
- Mortensen, H. (No Date). PhraseContext. Available at <http://www.hjkm.dk/>.  
[Accessed on 16 July 2009 22:43].
- Mujer Sana (2003). Healthy Women – Healthy Communities project.  
Available at <http://www.mujersana.ca/msproject/framework1-e.php>.  
[Accessed on 3 February 2009].
- O'Connor, S. (2003). Cheating and electronic plagiarism – scope, consequences and detection. Available at <http://www.caval.edu.au/research.html>. [Accessed on 7 June 2005 12:22].
- Pecorari, D. (2003). Good and original: Plagiarism and patchwriting

in academic second-language writing. Available at *Journal of Second Language Writing*.

*12 (2003) 317–345*

Pl@giarism (2008). Available at <http://people.few.eur.nl/span/Plagiarism/index.htm>.

[Accessed on 29 October 2009 19:20].

PlagiarismDetect (2009). Available at <http://www.plagiarismdetect.com/>. [Accessed on 25 October 2009 15:56].

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, Jan Svartvik (1974). *A Grammar of Contemporary English*. Longman.

Rawson, C [no date]. Breaking Down the Last Document Automation Barriers! Available at <http://www.infotivty.com/hwr.htm>. [Accessed on 12 September 2009 7:13 AM].

Rayson, P (2008). Wmatrix: a web-based corpus processing environment. *Computing Department, Lancaster University*. Available at <http://ucrel.lancs.ac.uk/wmatrix/>. [Accessed on 27 August 2009 19:51].

Roig, M (2006). Avoiding plagiarism, self-plagiarism, and other questionable writing practices: A guide to ethical writing. Available at <http://facpub.stjohns.edu/~roigm/plagiarism>. [Accessed on 19 August 2008 14:21].

Turnitin.com (2005a). *Rutgers University/Center for Academic Integrity*. Fast Facts and Stats on Cheating and Plagiarism. Available at [http://www.turnitin.com/static/products\\_services/latest\\_facts.html](http://www.turnitin.com/static/products_services/latest_facts.html). [Accessed on 30 May 2005 23:15].

Samoilovich S. (2009). Resources for English as a Second Language. Available at <http://www.usingenglish.com/>. [Accessed on 14 July 2009 18:28].

SFU (2007). Plagiarism Tutorial. Available at *Simon Fraser University*. <http://www.lib.sfu.ca/researchhelp/tutorials/interactive/plagiarism>. [Accessed on 20 August 2008 09:34 AM].

Stemler, S (2001). An Overview of Content Analysis. Practical Assessment, Research & Evaluation. Available at *ISSN 1531-7714. Yale University*.

<http://pareonline.net/getvn.asp?v=7&n=17>. [Accessed on 10 October 2007 14:02].

Swarthmore College. (2004a). Why is it a Problem? Available at <http://www.swarthmore.edu/NatSci/cpurri1/plagiarism/why.htm>. [Accessed on 30 May 2005 23:15].

Swarthmore College. (2004b). Deterrence Tips. Available at <http://www.swarthmore.edu/NatSci/cpurri1/plagiarism/deterrence.htm>. [Accessed on 30 May 2005 23:15].

Swarthmore College. (2004c). Detection. Available at <http://www.swarthmore.edu/NatSci/cpurri1/plagiarism/detection.htm>. [Accessed on 30 May 2005 23:15].

Thompson, S and Tony, O (2008). Plagiarism Prevention for Students. Available at *California State University*. <http://library.csusm.edu/plagiarism/>. [Accessed on 20 August 2008 09:34 AM].

TLAB (2009). Available at <http://www.tlab.it>. [Accessed on 27 August 2009 19:48].

Turnitin (2004a). Latest Facts. Previously available at [http://www.turnitin.com/static/products\\_services/latest\\_facts.html](http://www.turnitin.com/static/products_services/latest_facts.html). (Many other websites make reference to this page, which is no longer in existence) [Accessed on 30 May 2005 23:15].

Turnitin (2004b). How Plagiarism Prevention Works. Previously Available at [http://www.turnitin.com/static/products\\_services/process.html](http://www.turnitin.com/static/products_services/process.html). (Many other websites make reference to this page, which is no longer in existence) [Accessed on 30 May 2005 23:15].

Visualthesaurus.com (2009). Available at [www.Visualthesaurus.com](http://www.Visualthesaurus.com). [Accessed on 14 July 2009 9:45 AM].

Wapedia (2009). Available at [http://wapedia.mobi/en/Acoustic\\_phonetics](http://wapedia.mobi/en/Acoustic_phonetics). [Accessed on 9 April 2009 21:02].

Weeks, A (2006). Detecting plagiarism - Google could be the way forward. *British Medical Journal*. Volume 333, Page 706.

- WEKA (2009) Available at [http://www.cs.waikato.ac.nz/~ml/weka/gui\\_explorer.html](http://www.cs.waikato.ac.nz/~ml/weka/gui_explorer.html). [Accessed on 27 August 2009 19:44].
- Who's Who Among American High School Students [no date]. Previously Available at [http://www.turnitin.com/static/products\\_services/latest\\_facts.html](http://www.turnitin.com/static/products_services/latest_facts.html). [Accessed on 30 May 2005 23:15].
- WordNet (2006). Definition of Framework. Available at *Princeton University*. <http://wordnet.princeton.edu>. [Accessed on 30 September 2008 21:27].
- Wynne, M (2005). Stylistics: corpus approaches. Previously Available at *University of Birmingham*. [http://www.corpus.bham.ac.uk/conference2005/corpora\\_stylistics](http://www.corpus.bham.ac.uk/conference2005/corpora_stylistics). [Accessed on 14 August 2007].
- Zatyko, K (2007). Forensic Magazine – Defining Digital Forensics. Available at <http://www.forensicmag.com/>. [Accessed on 8 October 2008 15:05].

## *I n d e x*

- Academic Assignments, 25, 27  
Acquisition, 27, 28, 77, 139  
Analytic, 36  
Annotation, 32, 68, 69, 165  
Anti-Plagiarism, 21, 27, 45, 56, 63, 96, 157  
Antonymic, 35  
Audit, 36, 102, 155  
Authentic, 43, 77  
Author, 17, 43, 44, 45, 47, 50, 51, 52, 53, 54, 57, 67, 70, 72, 106, 109, 137, 138, 150  
Authorship, 20, 25, 33, 39, 43, 68, 70, 72, 92, 93, 99, 106, 110, 111, 113, 115, 116, 117, 119, 135, 136, 140, 149, 155, 157, 158, 162  
Bagging, 81  
Bagging And Tagging, 81  
Ballistics, 20, 73, 142  
Binary, 74  
Breach, 18, 24, 76, 141, 149, 156  
Calculus, 37  
Chain Of Custody, 76, 81, 146, 151  
Character Recognition, 94  
Class Characteristics, 104, 110  
Cognitive, 30  
Collusion, 52  
Compromise, 86  
Compromised, 28, 77, 81  
Computer, 17, 18, 20, 27, 28, 38, 68, 74, 75, 76, 77, 81, 84, 94, 96, 97, 115, 118, 142, 163, 164  
Computer Forensic Investigation, 29, 75  
Computer Security, 27  
Computerise, 38  
Conceptual Framework, 17, 18, 19, 21, 23, 24, 78, 83, 92, 140, 141, 143, 145, 147, 148, 149, 152, 155, 158, 159, 160  
Conclusions, 78, 83, 109, 112, 137, 147, 149, 154  
Concordance, 118, 119, 120, 121, 133, 139, 140, 143, 146, 157, 158, 163  
Constituent, 30, 31, 32, 33  
Constituents, 30, 32  
Copy, 27, 45, 77, 81, 151  
Copying, 27, 44, 47, 56, 67, 74, 77  
Core Constituent, 30, 31, 32, 33  
Corpus Analysis, 32, 136, 139, 157, 158  
Corpus Annotation, 69  
CRC, 75, 82, 98, 146, 155  
Crime, 28, 38, 41, 42, 56, 73, 74, 78, 79, 80, 81, 85, 86, 92, 110, 158  
Crimes, 27, 54, 74, 78, 80, 90  
Cryptomnesia, 44  
CTOSE, 73, 84  
Cultural, 67  
Cyber, 17, 18, 20, 21, 23, 24, 27, 28, 34, 37, 38, 74, 75, 80, 84, 97, 140, 141, 142, 147, 149, 152, 155, 158, 159, 165  
Cyber Forensic Auditing, 23, 24, 97, 140, 158, 159  
Cyber Forensics, 17, 18, 23, 24, 27, 28, 34, 75, 80, 97, 141, 142, 149, 152, 158, 159, 165  
Cybercrime, 38, 41, 42  
Cybernate, 38  
Data, 54, 77, 78, 143, 146, 151  
Database, 60, 96  
Definition, 18, 19, 93  
Design, 17, 25  
Digital, 19, 20, 24, 28, 29, 74, 77, 78, 79, 80, 82, 83, 84, 85, 86, 87, 88, 90, 146, 158  
Digital Forensics, 19, 28, 85, 86, 88, 90  
Digital Signatures, 77  
Discourse, 69, 105  
Discourse Analysis, 105  
DNA, 114, 140  
Duplicate, 53, 77, 78, 81  
E-Forensic Life Cycle, 73  
E-Forensics, 26, 27, 73  
Electronic, 20, 24, 27, 28, 29, 38, 41, 45, 65, 68, 73, 75, 77, 78, 85, 94, 97, 118, 142, 146, 151, 154, 155, 165  
Electronic Computer, 38  
Electronic Document, 28, 118  
Electronic Fraud, 20  
Electronic Material, 29  
Empirical, 18, 19, 25, 67, 149  
*Ethical Academic Writing*, 3, 155  
Ethical Motive, 40  
Ethical Writing, 17, 18, 19, 20, 23, 24, 25, 38, 40, 43, 44, 97, 156, 158, 159, 166  
Ethics, 23, 38, 40, 141, 149, 155  
Evidence, 17, 18, 27, 28, 44, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 85, 86, 87, 90, 91, 92, 106, 109, 110, 111, 118, 136, 146, 147, 151, 152, 155  
Executable Analysis, 29  
Extraction, 78, 80, 81, 82, 86  
Extrapolated, 43  
Findings, 78, 103, 137  
Forensic Auditing, 17, 18, 21, 33, 78, 84, 100, 118, 142, 149, 152, 154, 155  
Forensic Computing, 73  
Forensic Linguistics, 17, 20, 21, 23, 25, 99, 102, 107, 114, 142, 152, 155, 158, 165  
Forensic Pathology, 20, 73  
Forensic Psychology, 20  
Forensic Stylistics, 20  
Forensic Tools, 23, 24, 75, 156, 159

FORZA, 85, 87, 88, 89, 90, 143, 164  
 Framework, 17, 18, 21, 22, 23, 25, 57, 73, 75, 78, 79, 80, 83, 85, 88, 90, 92, 98, 140, 141, 143, 147, 148, 149, 152, 155, 158  
 Fraud, 19, 29, 42, 53  
 Frequencies, 43, 112, 114, 137, 151, 152  
 Google, 19, 65, 97, 167  
 Grammar, 25, 56, 99, 100, 111, 114  
 Guidelines, 20, 76  
 Hacker, 27  
 High Frequency, 70  
 Honourable, 40  
 Hypothesis, 47, 74, 83  
 Idiolect, 99, 111, 114, 116, 140  
 Image Analysis, 29  
 Imaging, 76, 77, 80, 81  
 Inadmissible, 75  
 Indeterminism, 34  
 Information, 28, 63, 69, 92, 155  
 Information System, 28  
 Institutional, 59  
 Intellectual Property, 19, 21, 23, 24, 27, 156  
 Intellectual Property Theft, 19, 21, 23, 24, 27, 156  
 Interface, 58  
 Internet, 20, 27, 45, 59, 65, 97  
 Interpretation, 43, 67, 104, 107  
 Interpretations, 34  
 Investigation, 17, 18, 19, 21, 24, 28, 29, 37, 67, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 90, 92, 98, 100, 102, 135, 136, 137, 143, 146, 147, 149, 150, 152, 155, 158, 159  
 Investigation Process, 29, 79, 82, 87, 88, 90, 143, 155  
 ISACA, 28, 75, 76, 77, 78, 164  
 Keywords, 82  
 Knowledge, 24, 46, 122  
 Law, 18, 21, 28, 41, 42, 53, 73, 74, 75, 76, 77, 81, 82, 84, 86, 90, 91, 107, 146  
 Law Breaking, 42  
 Legal, 21, 27, 73, 77, 81, 85, 87, 90, 91, 92, 105, 107, 139, 141, 143, 154, 158, 160, 164  
 Legal Requirements, 73  
 Lexeme, 35  
 Lexemes, 33, 35  
 Lexical, 29, 30, 32, 33, 34, 35, 69, 101, 137, 139, 155  
 Lexical Analysis, 43  
 Lexical Categories, 30, 32, 34, 35, 101, 139, 155  
 Lexical Patterns, 29  
 Linguist, 99, 111  
 Linguistic Analysis, 25, 103, 111, 140, 147, 148, 149, 152, 155, 158  
 Linguistics, 21, 26, 29, 67, 99, 101, 103, 109, 110, 142, 152, 158  
 Low Frequency, 43  
 mandate, 75, 136, 143, 146, 147, 150, 152, 155  
 MD5, 75, 82, 98, 146, 155  
 Microsoft, 60, 62, 63  
 Misconduct, 44  
 Misuse, 27, 84, 158  
 Morphology, 100  
 Network Analysis, 29  
 NHR, 95, 96  
 Node, 34  
 Nodes, 34  
 Norm, 116, 117, 136, 137, 155  
 Noun Phrase, 30, 31, 32, 33  
 Objectives, 17, 21, 23, 24, 25, 78, 90, 156  
 OCR, 28, 94, 136, 146, 151  
 Online, 63, 97  
 Online Tools, 25  
 Open Domain, 25, 67  
 Operating System, 29  
 Optical Character Recognition, 28  
 Paraphrasing, 43, 50, 51  
 Phonetics, 99, 100  
 Phonology, 100  
 Phrase Structure, 32, 101  
 Plagiarism, 17, 18, 19, 20, 21, 23, 24, 25, 27, 28, 29, 34, 43, 44, 45, 46, 47, 49, 50, 51, 52, 53, 54, 55, 56, 57, 60, 62, 63, 64, 65, 66, 67, 68, 73, 74, 94, 96, 97, 98, 115, 140, 149, 150, 151, 152, 154, 155, 156, 157, 159, 162, 164, 165, 166, 167  
 Pragmatics, 105  
 Pre-Determiners, 31  
 Pre-Empirical, 18, 25  
 Presentation, 18, 69, 76, 78, 79, 80, 91, 92  
 Principals, 17, 20, 86, 149  
 Probability, 75, 93, 99, 137, 138  
 Problem-Solution Oriented Research, 21, 22, 147, 148, 149  
 Pronoun, 30, 31, 32, 72  
 Propositional Phrase, 30  
 Prosecution, 73, 74, 82  
 QD, 115  
 Qualitative, 25, 106, 138, 146  
 Quality, 18, 23, 24, 68, 72, 76, 115, 156, 159  
 Questioned Document, 115  
 RAID, 29  
 Recommendations, 78, 154  
 Reference, 44, 45, 50, 57, 84, 110, 136  
 Relationship, 18, 19, 20, 23, 24, 35, 74, 83, 114, 143, 152  
 Report, 78, 137, 138, 147, 152  
 Research Questions, 17, 21, 24, 97, 136, 146, 150, 153, 154, 156, 159  
 Researcher, 21, 24, 72, 92, 93, 94, 155, 156, 159  
 Role Players, 78, 87, 88, 150



Scan, 94  
 Science Direct, 19  
 Scope, 23, 57, 73, 75, 78, 87, 136, 165  
 Search And Seizure, 81  
 Self-Plagiarism, 53, 54, 166  
 Semantic, 69  
 Semantics, 100, 105  
 Semiosis, 34  
 Sentence Patterns, 29, 30, 33, 99  
 Software, 57, 98, 115, 118, 121, 146  
 Software Programs, 25, 78, 120, 146  
 Spider Diagram, 34, 42  
 Spider Diagrams, 34, 42  
 Spider-Grams, 35  
 Statistical, 116, 138  
 Strategy, 21, 87  
 Stylistic Analysis, 27, 67, 70, 73, 115, 155  
 Stylistics, 18, 20, 23, 24, 25, 26, 67, 68, 73, 94,  
 99, 106, 114, 115, 116, 140, 165, 168  
 Subproblem, 23  
 Synonymic, 35  
 Syntactic, 21, 30, 99, 155  
 Syntactic Patterns, 30, 155  
 Syntax, 25, 29, 99, 100, 137  
 System, 64  
 Tagging, 69, 81  
 The Lexicon, 100  
 Tools, 73, 84, 97, 98, 118, 121  
 Tools And Techniques, 21, 23, 156  
 Transitive Sentences, 30, 101  
 Tree Diagram, 30, 31, 33  
 Unethical, 18, 19, 20, 40, 57, 86  
 Unethical Academic Writing, 18, 19  
 Unethical Writing, 18, 19, 20  
 Unintentional, 44  
 Validity And Reliability, 28  
 Verb Phrase, 30  
 Video Analysis, 29  
 Violate, 21, 74  
 Word Frequency, 25, 43, 121, 152, 155  
 Word Frequency Analysis, 25, 43, 121, 155  
 Writing Style, 43, 51, 57, 67, 102, 111, 112,  
 139  
 Zachmans Framework, 88

**TO WHOM IT MAY CONCERN**

7 May 2010

This dissertation, entitled *Forensic computing strategies for ethical academic writing*, by Sashen Govender, has been edited to ensure technically accurate and contextually appropriate use of language.

Sincerely

A handwritten signature in blue ink that reads "CM Israel". The signature is written in a cursive style with a long horizontal stroke at the bottom.

**CM ISRAEL**  
**BA Hons (UDW) MA (UND) MA (US) PhD (UNH)**  
**Language Editor**

# Volsum, Chetty & Lax

*Attorneys, Conveyancers And Administrators Of Estates*

Director:

Kooben Chetty B.Proc. (Natal)

Professionally assisted by:

Pregila Chetty BA.LLB. (UDW)

Anisha Ramrathan B.Soc.Sc. LLB. LLM. (Natal)

4 George Street  
Pietermaritzburg, 3201

P. O.Box 8218  
Cumberwood, 3235

Dx : 62 Pmb  
E-mail : anishavcl@iafrica.com  
Tel : (033) 3948115  
Fax : (033) 3948150

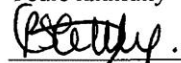
---

## TO WHOM IT MAY CONCERN

10 June 2010

It is my opinion that the legal aspects in Sashen Govender's dissertation, entitled *Reading between the lines: Forensic computing strategies for ethical academic writing*, has been adequately presented.

Yours faithfully



\_\_\_\_\_  
Pregila Chetty  
VOLSUM CHETTY & LAX

\_\_\_\_\_  
PROP: Volsum Chetty Inc.  
2000/025790/21



RESEARCH OFFICE (GOVAN MBEKI CENTRE)  
WESTVILLE CAMPUS  
TELEPHONE NO.: 031 – 2603587  
EMAIL : sshrec@ukzn.ac.za

---

23 JUNE 2010

MR. S GOVENDER (204002155)  
IS & T

Dear Mr. Govender

PROTOCOL REFERENCE NUMBER: HSS/0618/08M  
NEW PROJECT TITLE: Forensic Computing Strategies for Ethical Academic Writing

**APPROVAL AND CHANGE OF DISSERTATION TITLE**

I wish to confirm that ethical clearance has been granted full approval for the above mentioned project:

Any alteration/s to the approved research protocol i.e. Questionnaire/Interview Schedule, Informed Consent Form, Title of the Project, Location of the Study, Research Approach/Methods must be reviewed and approved through an amendment /modification prior to its implementation. In case you have further queries, please quote the above reference number. PLEASE NOTE: Research data should be securely stored in the school/department for a period of 5 years

Best wishes for the successful completion of your research protocol.

Yours faithfully

.....  
PROFESSOR STEVEN COLLINGS (CHAIR)  
HUMANITIES & SOCIAL SCIENCES RESEARCH ETHICS COMMITTEE

cc. Supervisor (Prof. R Klopper)  
cc. Mrs. C Haddon