The Case for Cyber Forensic Linguistics

Rembrandt Klopper

Abstract

This contribution focuses on one of the most vexing problems facing present-day higher education, namely how to identify, curtail and eliminate plagiarism among students and academics alike according to robust forensic audit protocols that do not leave institutions open to subsequent legal challenges. Academia is faced with the stark choice between taking a leap of faith by trusting software programs that algorithmically establish superficial similarities between text fragments on non-semantic grounds, or by using an array of text analysis procedures that scientifically determine degrees of similarity between texts on rational grounds by employing proven principles of linguistic and text analysis. This article presents a multi disciplinary conceptual framework that integrates Forensic Computing and Forensic Linguistics into a comprehensive forensic auditing framework within a mother interdiscipline like Informatics, Information Systems & Technology or Computing. I specifically propose the formal recognition of a new interdiscipline within the field of forensic analysis, namely *Cyber Forensic Linguistics*. This nascent area of learning can be created by identifying and integrating the analytical procedures used in disciplines like Linguistics, Informatics and Auditing into a coherent curriculum that will go a long way to combat plagiarism on semantic grounds, and help promote ethical academic writing.

Keywords: Anaphora, categorisation, computer forensics, cyber forensics, computational forensics, conceptual framework, deep web, ethical academic

writing, ethics, forensic audit, forensic linguistics, internet, invisible web, lexemes, phoric referencing, plagiarism, problem-solution oriented research, semantic roles, surface web, word frequency analysis, world wide web.

A Problem-Solution Oriented Research Approach

This contribution implements a problem-solution oriented approach to research that was developed as a dissertation writing framework in Klopper (2008). It forms the basis of the procedure that is proposed for cyber forensic audits of electronic documents where authorship is at issue.

A problem-solution oriented approach implements the procedure of assessing the extent to which one is able to answer research questions that have been derived from research problems. In order to identify a problem that causes a system to malfunction the researcher first has to inferentially diagnose the problem and then envisage how the system should be when functioning optimally. Finally, the researcher has to detail how the system should be reconfigured so that it would function optimally. The essence of the problem-solution oriented approach therefore is a bifocal perspective that alternates between problematic and ideal versions of the system under scrutiny.

The Problem-set under Investigation

The alignment matrix given as Table 1 below summarises the problems identified and their associated research questions:

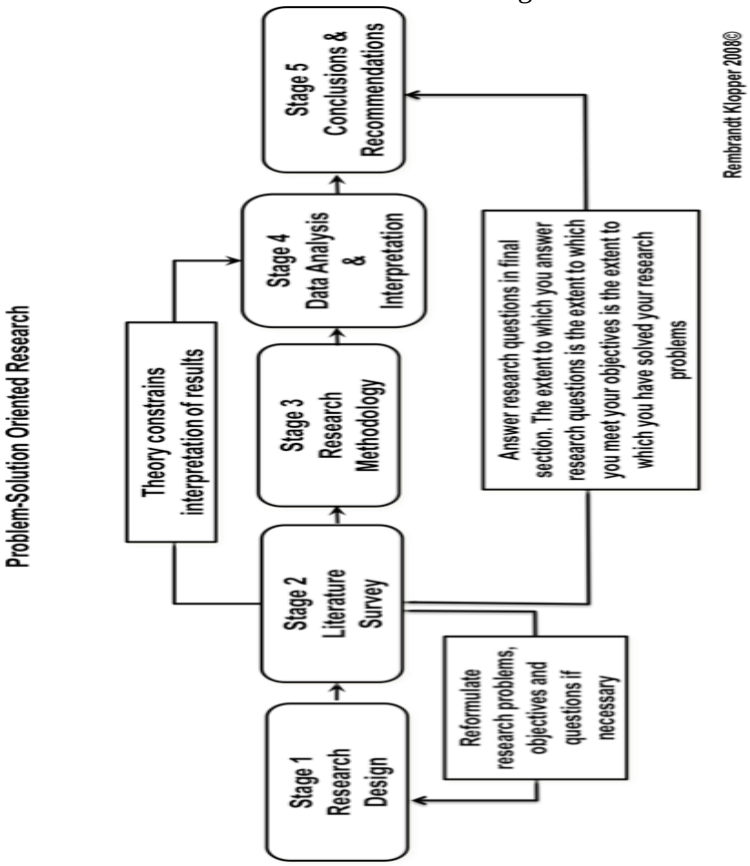
Table 1: Problem-research question alignment matrix

General Problem	**Subproblems**	**Research Questions**
Academics do not yet understand the extent and sources of plagiarism in academia.	1. The sources of plagiarised academic material have not yet been identified. 2. The resources for tracking and eliminating plagiarism have not yet been identified. 3. It has not yet been determined how methods of text analysis	1. Where on the Internet do plagiarists obtain plagiarised sources? 2. What resources can be used to track and eliminate plagiarism? 3. What methods of

	could serve as forensic tools to assess the quality of ethical academic writing.	text analysis could serve as forensic tools to identify plagiarism on linguistic grounds?

The problem-solution oriented approach to research is outlined in Figure 1:

Figure 1: Problem-solution oriented research design



At the most basic level the prevention of plagiarism is a problem of an *epistemic* nature, namely ignorance (lack of knowledge) among members of the academic community regarding the full extent of electronic information directly available for download on the Internet (the so called surface web), let alone data only indirectly available for download on the deep web/ invisible web with the aid of specialist online resources or services (Bergman (2009), Wright (2009)). The concepts 'surface web' and 'deep web' are briefly discussed in a subsequent section.

It should also be acknowledged that the distinction between learning from others, and pretending that one is the author of a particular idea or set of ideas isn't always straightforward. Due to the way in which humans think and learn it isn't a simple matter to identify the dividing line between genuine research and mere regurgitation. Learning in general, and scientific advances in particular, are collective enterprises that are deeply analogical in nature – what we know we have learnt from others. We learn to look beyond the horizon of present-day knowledge by standing on the shoulders of those who set out to mend gaps in our knowledge before we did. In order not to trivialise and debase our disciplines while under pressure to 'publish or perish' our only hope is to conduct research guided by ethical principles, by identifying real problems that need to be solved, to learn from our peers by consulting and acknowledging the solutions that they have proposed to the problems that we identified, by critically engaging with their solutions—and finally, by being willing to have one's own solutions critically assessed by one's peers. This contribution proposes, as a measure of last resort, a conceptual framework to implement cyber forensic auditing as the sharp end of critical engagement in order to distinguish between genuine research and mere regurgitation.

The weakness of anti-plagiarism programs is that they determine similarities between text fragments on non-semantic grounds and that they only compare documents of questionable authorship with documents that are directly accessible on the surface web rather than also accessing documents that are only indirectly accessible on the part of the Internet known as the deep web.

Anti-plagiarism programs like MyDropBox and TurnitIn are therefore mainly useful to identify instances of crass plagiarism because they function on non-semantic grounds, often misidentifying text fragments like

'with his hat on' or 'most Sundays' as having been copied, often identifying multiple sources for it on the internet. It is also not generally appreciated in academic communities that students, or their paid *agents provocateurs*, use such programs to polish up poorly referenced documents lifted from the surface and deep web by merely referencing the sources 'identified' by the programs.

The Nature of Conceptual Framework Design

Because this contribution proposes a conceptual framework for conducting a forensic audit of documents to determine plagiarism, it is appropriate that the nature of conceptual frameworks be briefly explained. Also, because the process of conceptual framework design is not commonly used in research, I briefly characterise it and its role during the research process.

The term 'conceptual framework' refers to a pre-empirical procedure that is employed during the planning phase of research in emerging fields of research where little or no prior research has been done. A conceptual framework interrelates all relevant factors that need to be brought to bear during analysis to explain why something is being done in a particular way, or needs to be done in that way. It is devised to guide analysis. If one for instance wanted to perform content analysis of a document, one would first devise an analytical framework of all categories for which one would systematically identify exemplars throughout the document, as well as a system of codes that one would employ to tag each occurrence of exemplars that belong to the different categories, and finally a quantification procedure to calculate the relative occurrences of members of each category. Therefore, conceptual frameworks first have to be designed to direct and constrain analysis, before actual analysis can be performed.

Only four definitions have been found on the Internet for the term 'conceptual framework':

(1)

- 'A conceptual framework is used in research to outline possible courses of action or to present a preferred approach to an idea or thought' (Wikipedia 2009, Conceptual Framework).

(2)

- 'The Conceptual Framework is a linked set of agencies that assist in the critical analysis of art. They can be discussed as separate entities or linking to the Frames (cultural, structural, subjective, and postmodern). The four agencies are enumerated as follows:
- The Role of the Artist (Artist)
- The Roles and Values of the Audience (Audience)
- Artworks As Real Objects (Artwork)
- How Interests in the World Are Represented (World)' (Wikipedia 2009, The Conceptual Framework).

(3)

- 'Because of the richness and complexity of the material, coverage [of the term conceptual framework] has been planned in accord with the following broad categories or subject areas, each of which has been further subdivided by particular topics' (Hattendorf 2005).

(4)

- 'A [conceptual framework is a] theoretical model devised to establish a series of conceptually sound practical accounting rules' (Pearson Education 2004).

According to the above definitions conceptual framework design is employed in academic protocols as diverse as research design, critical analysis of art, maritime studies (Hattendorf 2005) and the formulation of accounting rules (Pearson Education 2004). These definitions also reveal that conceptual frameworks are used to outline *possible courses of action*, or *preferred procedures*, that they can be used to detail *agency role relationships* relevant to particular fields of inquiry, and that they could stipulate conceptually sound *procedural ('accounting') rules*. All of the before-mentioned approaches have in common that they are forms of qualitative research that systematically focus on meaningful relationships among the elements of the entity that is being studied.

This synthesis of the nature and function of conceptual frameworks that I have presented here is taken as point of departure for designing the conceptual framework that I am proposing for integrating electronic forensics

and forensic linguistics into a comprehensive framework of forensic auditing in order to curtail plagiarism and promote sound academic writing.

To summarise, the conceptual framework that is proposed here implements:

1. The best-practice procedure for forensic auditing,
2. The use of agency role relationships to analyse interactions between the parties involved in events like authoring, replicating, copying and pasting information in written documents that require the forensic audit, and
3. Conceptually sound procedural rules for forensic auditing.

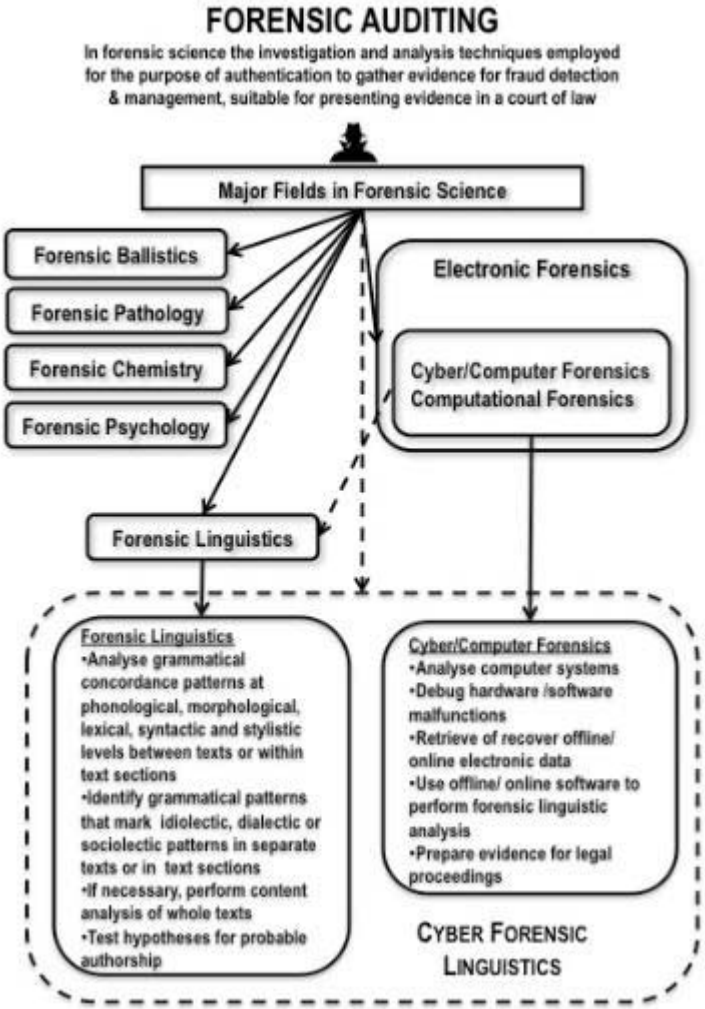
Computer Forensics and Computational Forensics

In this section I provide clarifications for the terms 'computer forensics' and 'computational forensics,' after which I point out a major shortcoming of these two approaches.

- 'Computer forensics is a branch of forensic science pertaining to legal evidence found in computers and digital storage media. Computer forensics is also known as digital forensics' (Wikipedia 2009).
- 'Computational forensics is a quantitative approach to the methodology of the forensic sciences. It involves computer-based modeling, computer simulation, analysis, and recognition in studying and solving problems posed in various forensic disciplines. CF integrates expertise from computational science and forensic sciences.' Computational forensics is used to analyse objects, substances and processes that are investigated, mainly based on pattern evidence, such as toolmarks, fingerprints, shoeprints, documents etc., but also physiological and behavioral patterns, DNA, digital evidence and investigation' (Wikipedia 2009).
- In Figure 2 below I list the major fields or forensic science, and I indicate how the procedural integration of two of those fields – electronic forensics and forensic linguistics – could create an

emerging new field of forensic science, namely *cyber forensic linguistics*.

Figure 2: Proposed field of cyber forensic linguistics among major fields of forensic science on which forensic auditing draws



From the point of view of using computer forensics and digital forensics to identify and counteract plagiarism, the major shortcoming with the above-mentioned two approaches clearly is that forensic auditors who depend on these techniques tend to restrict electronic evidence gathering and analysis to individual machines or local area networks, whereas the crime scene is distributed to wherever the boundless resources of the Internet are misused during plagiarism. It is therefore self-evident that the scope of cyber forensics/ computer forensics/ digital forensics and computational forensics should be widened to include remote data search, retrieval and analysis protocols.

Because the protocols for cyber forensics and computational forensics are well documented (e.g., Marcella and Greenfield (2006), ISACA (2004), Forensic Magazine (2009) and Carrier and Spafford (2004)), but those for forensic linguistics are less well documented, and because linguistic analytical procedures are central to the establishment of the likelihood of authorship in electronic documents during forensic audits, I will briefly outline the main aspects of forensic linguistics in the next section, followed by a brief explanation of the forensic audit process.

Anti-plagiarism software and services on the Internet

Although I have done a fair amount of research about the online and downloadable electronic resources available for cyber forensic auditing, I am identifying resourcess in this contribution without assessing their usefulness, to ensure that the focus of the case that I am presenting for cyber forensic auditing does not become unfocussed. I am identifying software programs (referenced in the bibliography) besides MyDropBox and TurnitIn that work on similar principles, and that can therefore also be used to help identify instances of Internet plagiarism on the surface web (URLs provided in the References section):

1. Article Checker
2. CopyCatchGold
3. CopyScape
4. Dupli Checker
5. Eve2
6. Glatt Plagiarism Screening Program
7. Google as an Anti-Plagiarism Tool

8. Jplag
9. Pl@giarism
10. PlagiarismDetect
11. SafeAssign

Deep Web and Shallow Web Search Engines

While Internet users assume that they *directly* retrieve data from servers on the Internet when making search engine queries, they actually access historical data which has been pre-retrieved from accessible areas of the internet by search-engine-based programs known as crawlers, spiders or bots, that constantly scour the internet for new data which is then saved on the mainframe servers of the search engine companies. Therefore, when one uses a particular search engine, one only *directly* accesses the pre-retrieved data that relate to one's query from the mainframes of the search engine companies. It is also not generally appreciated that such crawlers can only access the *shallow web*, the relatively limited part of the internet where data is *directly* available via search engines, and that search engine crawlers cannot retrieve and save information that is *indirectly* available on the part of the internet that is know as the *deep web*.

According to Bergman (2009) public information *indirectly* available on the deep web is 400 to 550 times larger than information *directly* accessible on the Internet (the world wide web), consisting of an estimated 550 billion individual documents, amounting to 7,500 terabytes of information (compared to 19 terabytes on the Surface web). Bergman also states that on average the deep web receives about 50% greater monthly traffic than surface sites, that deep web sites tend to be narrower with deeper content than surface web sites, with nearly half of deep web content residing in topic-specific databases, and finally that nearly 95% of the deep web contains publicly accessible information for those who know how to access it.

Meta-search engines simultaneously search the shallow web via other search engines to broaden the spectrum of information retrieved. As indicated, the internet is divided into the *surface web*, referring to homepages and databases that are directly accessible via search engines and meta-search engines, and the *deep web* (sometimes also referred to as the invisible web),

referring to homepages and databases that are only indirectly accessible. The resources listed below have been proven to be effective to access the deep web:

- **SurfWax**: This meta-search engine searches the Deep Web.
- **Academic Index**: This meta-search engine only searches databases and resources that have been approved by librarians and educators, including repositories for online dissertations and other scholarly documents that are available for download.
- **Clusty**: Clusty simultaneously searches multiple search engines, then clusters the search results, thereby making visible information that would normally be hidden far back in search results due to low ranking.
- **Turbo 10**.: This meta-search engine has specifically been designed to search the Deep Web.
- **World Curry Guide**: This meta-search engine has a strong European orientation.

Semantic Search Engines

Semantic search engines emulate the way the human brain categorises information to ensure more relevant search results. Prominent semantic search engines are:

- **Hakia:** This semantic search engine only accesses websites that have been recommended by librarians typically websites that contain online dissertations and scholarly documents.
- **Zotero:** This is an add-on for the Firefox browser that helps one to organise research material by collecting, managing, and citing references from material downloaded from the Internet.
- **Freebase:** This is a community-powered database that covers information on millions of topics.
- **Powerset:** One can enter a topic, phrase, or question in the search slot to find information from Wikipedia with this semantic application.

- **Kartoo:** One can enter any keyword in the search slot to receive a visual concept map of information available on the Internet that pertain to that keyword. By hovering one's mouse pointer over each information node one gets a thumbnail of the website where the data is located.
- **DBpedia:** This another Wikipedia-specific resource, allowing one to ask complex questions to get answers from Wikipedia.
- **Quintura:** When one enters a search term in the search slot the web page splits into two panels. On the left hand side a 'cloud' of related terms appear, and on the right hand panel links appear for the search term. When one hovers one's mouse pointer over a related term or phrase an entirely different list of links, appropriate for the new term, appears on the right hand panel.
- **Evri:** This search engine provides one with results across a range of media, from articles, papers, blogs, images, audio, and video on the Internet.
- **Gnod:** When one repeatedly uses this search engine to search for information about books, music, movies and people on this search engine, it remembers your interests and focuses the search results in that direction.

The Concept 'Ethical Academic Writing'

Ethical academic writing clearly forms part of ethical behaviour. In this section I focus attention on the concepts 'ethical behaviour' and 'moral behaviour' in view of the fact that a clear distinction is often not drawen between the terms 'ethics' and 'morality' in scientific and philosophical literature. The first facts that I want to report is that only one definition has been foud for 'ethical behaviour' on the Internet, and that no definitions have been found for 'moral behaviour'.

- **Moral:** '(concerned with principles of right and wrong or conforming to standards of behavior and character based on those principles) *'moral sense'; 'a moral scrutiny'; 'a moral lesson'; 'a moral quandary'; 'moral convictions'; 'a moral life'* (WordNet Search—3.0).

- **Moral Philosophy:** 'Moral philosophy is the area of philosophy concerned with theories of ethics, with how we ought to live our lives' (Moral Philosophy .Info).
- **Ethical behaviour:** 'Ethical behaviour is behavior that conforms to accepted professional standards of conduct' (Donneley (2005)).

A comparison of the above definitions shows that no clear distinction is made in academic and philosophical circles between concepts that relate to personal and professional norms for ethical and moral behaviour, including norms for ownership of intellectual property. Such norms of intellectual ownership form a crucial part of a forensic auditing framework to curtail plagiarism and promote ethical academic writing.

Forensic Linguistics

Non-linguists tend to underestimate how complex human language is because we are neurologically hardwired to develop innate language capacity from shortly after birth. Whatever languages one uses to think and communicate, one does so subconsciously. The complexity of language as a symbolic system has led to the establishment of a variety of schools of linguistics over the past two hundred years. Involving Linguistics in forensic analysis therefore presents a particular challenge, because of the phletora of schools of thought, such as Descriptive Linguistics, Lexical-Functional Linguistics, Applied Linguistics, Stylistics, Relational Grammar, Generative Grammar, Socio-linguistics, Psycholinguistics, Computational Linguistics, Corpus Linguistics and Cognitive Linguistics, to name but a few. Whatever system of linguistic analysis one utilises as part of forensic linguistics, one analyses the structure of language at the phonological (sound pattern) level, the morphological (word structure) level, the syntactic (phrase and clause structure) levels, the semantic level and the stylistic level. These aspects of linguistics will be briefly attended to in the section that follows.

A Short and Intense Introduction to Linguistics

This section synthesises particular aspects of linguistics – words and their meanings, and the relationship between such words and other words in

sentences—that reveal inferential relationships between language forms that enable one to make a nuanced analysis of texts to help establish authorship. McMenamin *et al.,* (2002) provides a sound introduction to the basic aspects of Forensic Linguistics, focusing on particular aspects of linguistics relevant to Forensic Linguistics, like the areas of Applied Linguistics, Stylistics and Linguistic Variation that are relevant to Forensic Linguistics. The additional primary references consulted for this summary are Quirk *et al.,* (1974), Sampson 1980), Johnson (1987), Lamb (1998), Klopper (1999, 2003), Taylor (2003), Ungerer and Schmid (2006), Holcombe (2007), Baldwin *et al.* (no date), Baldwin *et al.* (2009) and Haas (no date). These texts describe grammatical patterns from the highly compatible perspectives of Lexical-Functional Linguistics, Cognitive Linguistics, Computational Linguistics and Corpus Linguistics.

In view of the fact that this project focuses on written language, various aspects of spoken language (sound patterns, also known as phonology, the importance of analysing pronunciation of words in particular regional dialects, degrees of emphasis on particular syllables of polisyllabic words and tone of voice) are acknowledged as crucial aspects of general linguistics, but are nonetheless excluded from this contribution.

The information that is provided about Linguistics here is done with the objective of clearly demonstrating the level of complexity of human language as a symbolic system, in order to emphasize that any course in Cyber Forensic Linguistics should impart in-depth knowledge of language patterns, and that a Linguistics-light approach would doom the project to failure from the outset.

From Concept Conflation to Word Formation

Humans build up concepts about things around them by combining basic image schemas like point, line, centre, periphery, circle, square, triangle, long, short, horizontal, vertical, diagonal, close proximity, distant proximity, in front of, behind, smooth, coarse, regular, irregular, move, rest, source, route/path, target etc. These image schemas emanate from our physiological makeup and our vertical orientation when we are active (Johnson 1987). By combining basic image schemas into complex image schemas humans derive dynamic concepts. If one for instance combines the image schemas long and

vertical in relation to an entity that is perceived as a single whole, one conceptualises the word *tall*. By combining the image schemas centre, periphery, vertical, move, regular sequence, same direction and balance one conceptualises words like *rotate, spin* and *pirouette*.

The Fundamental Role of Categorisation in Word Formation

Categorisation forms a crucial part of humanity's language capacity. Humans use the shared and differentiating attributes of entities in inter-linked neural pathways in the brain (Lamb 1998) to categorise things, processes and events. We discern attributes such as compact, diffuse, round, square, smooth, coarse, bright, dull, dark, sweet, sour etc. by means of our senses (sight, hearing, taste, smell and feeling), and we use such attributes to categorise and inter-link words.

Humans categorise entities on at least three levels, namely a superordinate level (*plant*), a basic level (*tree*) and a subordinate level (*oak tree*). Ordinary people have pictographic gestalts for entities at the basic level, but not at the superordinate and the subordinate levels. We can doodle basic level images of a *man* or a *woman*, but only an artist can 'doodle' posture-rich subordinate level human figures that suggest specific body postures or body movements as in Figure 3:

Figure 3: Basic level images of a man and a woman (Webdings font set), followed by subordinate level, detail-rich images of humans (Ihminen font set).

According to Mandler (2004), babies develop mental problem-solving skills like the ability to compare and categorise entities and the ability to distinguish between purposeful, goal-directed behaviour from non-purposeful behaviour between six and nine months old, clearly before they actually develop language-specific competence and they learn basic level words before they learn superordinate or subordinate level words Markman (1990). They for instance learn *cat* and *dog* before they learn *animal* or

Siamese cat and *bull terrier*. Basic level words tend to form part of figurative language more often than superordinate level or subordinate level words. Humans for instance represent their ancestry by means of *family trees*, not by means of *family plants* or *family oak trees*.

Prototypical and Atypical Members of Lexical Categories

Categories have prototypical members that share all major attributes, and atypical members that share only some attributes (Taylor (2003) and Ungerer & Schmid (2006)). *Mammals* for instance are warm-blooded, hairy, earth-bound animals with four limbs and teeth, whose offspring are born alive from the female member of their species, and who obtain nourishment by suckling her mammary glands during the early stages of their existence. By this definition *antelope, lions, rhinos, wolves* and *humans* are prototypical mammals, but *bats* are not because they are flying mammals, nor are *beavers, seals, walruses, whales, dolphins,* or *manatees* because they are aquatic mammals.

Meaningful language symbols are known as lexemes. Lexemes are associated with particular grammatical categories. A word like *crazy* is a prototypical adjective, while a set expression like *off his rocker* is an atypical one. *Atypical lexemes are more useful than prototypical ones during text analysis to attribute likely authorship.*

The Conflation of Concepts in the Formation of Lexemes

Words are associated with lexical categories like nouns (*woman*), verbs (*break*), adjectives (*tall*), adverbs (*forwards*), pronouns (*you*) and prepositions (*on*).

Each lexeme consists of a number of concepts that are conflated (combined) and associated with a particular sequence of speech sounds or written letters. The sequence of speech sounds [m]+[æ]+[n], and the letters *m+a+n* are neurologically associated with the concepts [living, male, adult member of the human species]. Lamb's relational network theory of language, presented in Lamb (1998), provides a credible explanation for how the production of speech sounds and the formation of concepts are neurologically interrelated.

Conceiving Entities and Events in Schematic and Detailed Formats

Humans have the ability to conceptualise entities in *schematic* (very general) or in *detailed* terms. One can for instance say: *There is something under the table*, or *There is a little red metal toy motor car with plastic windows and scratched bumpers under the table.*

All elements of language, including grammatical structures, are *symbolic* (meaningful). Particular grammatical structures have *schematic* (general) background meanings that give them the valence (binding potential) to accommodate particular words.

Sentence Patterns: Phrases and Clauses, Topic and Focus

In *transitive* clauses for instance the major grammatical structures have the following schematic meanings: *Subject*: the party that supplies the energy for an interaction and that actively controls the course of an event which affects another party. *Transitive verb*: portrays an interactive event during which energy is transferred from an active, controlling party to a passive, affected party. *Object*: the passive party that is controlled and affected by the actions of an active party. Because a transitive clause has the above-mentioned schematic meanings it casts a semantic shadow that enables it to accommodate a great variety of utterances that portray interactions, like for example: *The dog is chasing the cat, my aunt is writing a novel, James cooked dinner last night* and *the boy is slicing the salami.*

Sentences are technically referred to as 'clauses' in most syntax textbooks, and portray real-world interactions in the course of events. More precisely, they portray the interactions between entities during actual or envisaged (hypothetical) events.

Sentences are made up from two types of clauses, namely independent clauses as in [**INDEPENDENT CLAUSE** *The child is shivering*], and dependant clauses that are subordinate to independent clauses as in [INDEPENDENT CLAUSE *The child is shivering,* [**DEPENDENT CLAUSE** *because he is feeling cold*]], or the less common [[**DEPENDENT CLAUSE** *Because he is feeling cold*] **INDEPENDENT CLAUSE** *the child is shivering*]]. A clause contains only one verb, but can contain a number of nouns such as in [**SENTENCE** [**NOUN 1** *the dog*] [**VERB** *chased*] [**NOUN 2** *the cat*] *up* [**NOUN 3** *the tree*] [**NOUN 4** *this morning*]].

Sentences are organised on a binary pattern according to which the first phrase of the clause occupies the privileged Topic position, with the rest of the sentence occupying the Focus position, the new information that is being provided about the topic under consideration. In transitive clauses the subject noun phrase occupies the Topic position and the verb phrase the Focus position as in [$^{\text{SENTENCE}}$ [$^{\text{TOPIC}}$ [$^{\text{SUBJ NP}}$ *the dog*] [$^{\text{FOCUS}}$ [$^{\text{VERB PHRASE}}$ [$^{\text{VERB}}$ *chased*] [$^{\text{OBJ NP}}$ *the cat*] *up* [$^{\text{PLACE NP}}$ *the tree*] [$^{\text{TIME NP}}$ *this morning*]].

Topicalisation is the process according to which sentence elements that are not normally used in the Topic position, are promoted to the Topic position for special attention as in [$^{\text{SENTENCE}}$ [$^{\text{TOPIC}}$ [$^{\text{TIME NP}}$ *this morning*] [$^{\text{FOCUS}}$ [$^{\text{SUBJ NP}}$ *the dog*] [$^{\text{VERB PHRASE}}$ [$^{\text{VERB}}$ *chased*] [$^{\text{OBJ NP}}$ *the cat*] *up* [$^{\text{PLACE NP}}$ *the tree*]], or as in the passive construction [$^{\text{SENTENCE}}$ [$^{\text{TOPIC}}$

[$^{\text{OBJ NP}}$ *the cat*]] [$^{\text{FOCUS}}$ [$^{\text{VERB PHRASE}}$ [$^{\text{VERB}}$ *was chased*] *up* [$^{\text{PLACE NP}}$ *the tree*] *by* [$^{\text{SUBJ NP}}$ *the dog*] [$^{\text{TIME NP}}$ *this morning*]].

Phrases essentially are lexical categories like noun, verb, article, adjective and preposition that are hierarchically linked to provide a coherent brief account of an event. Examples are given of two prepositional phrases, namely *with a knife,* which encodes instrumentality and *for what he believed in,* which encodes the reason for the event, respectively shown in Figures 4 and 5:

Figure 4: English prepositional phrase encoding instrumentality

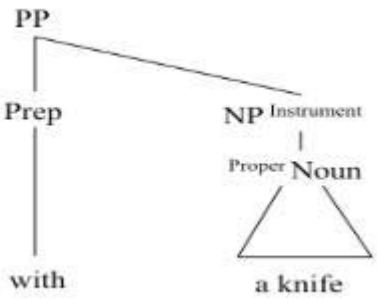
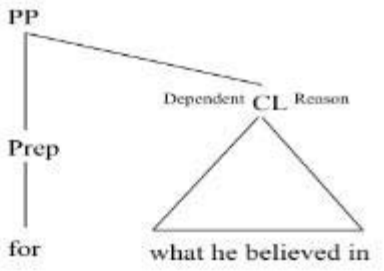
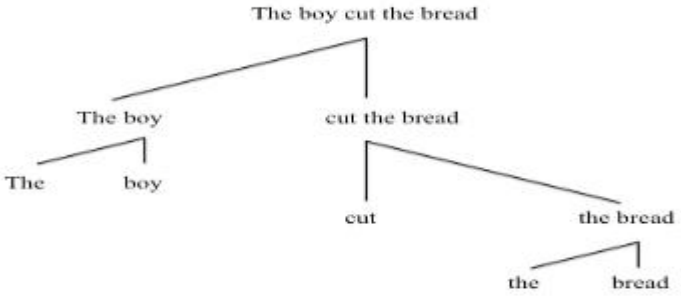


Figure 5: English prepositional phrase containing a dependant clause encoding the reason for the portrayed event



One of the basic types of clauses is the transitive sentence that hierarchically links a subject noun phrase with a verb phrase, which in turn requires an object noun phrase as complement. This is known as the SVO word order of transitive sentences that encode transitive cause-and-effect events. In such transitive constructions the entity stipulated by the subject noun phrase provides the energy for the event stipulated by the transitive verb and controls the course of the event, while the entity stipulated by the object noun phrase is passively involved, absorbs the energy and is affected by it as in *the boy cut the bread* in Figure 6:

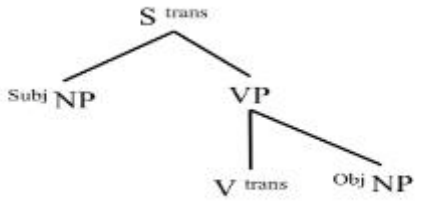
Figure 6: The hierarchic organisation of a typical English transitive clause



The formal notation for such transitive clauses is given below, showing that a transitive sentence (S^{trans}) consists of a subject noun phrase

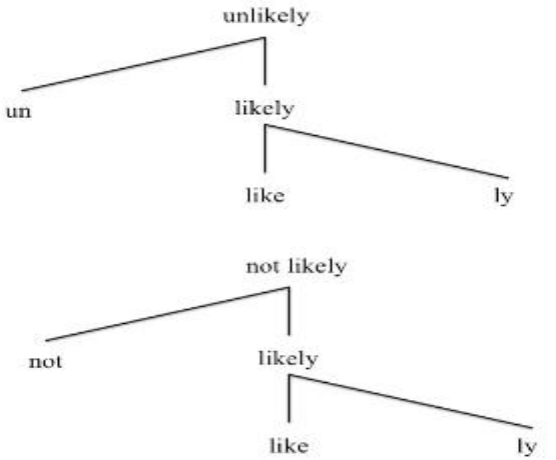
($^{\textbf{Subj}}$ NP) and a verb phrase (VP), which in turn consists of a transitive verb (V $^{\textbf{trans}}$) and an object noun phrase (($^{\textbf{Obj}}$ NP), as shown in Figure 7:

Figure 7: The hierarchic organisation of English transitive sentences



Morphological constructions (complex lexemes built up from smaller lexical units known as 'morphemes') are organised on the same hierarchical principles as clauses as can be seen in Figure 8:

Figure 8: The hierarchic organisation of the English complex lexeme 'unlikely'



From Structure to Meaning: The Distribution of Semantic Role Relationships within Sentences

In order to distinguish nouns from one another in a clause one conceptually assigns semantic roles like Agent, Patient, Experiencer and Stimulus to them. Other semantic roles used to identify entities during interactions are Source, Target, Path, Goal, Benefactor, Beneficiary, Instrument, Locus/Place and Time.

During conversation participants automatically assign semantic roles to the nouns in a clause like *the dog chased the cat up the tree this morning*, which will give the following representation: [$^{\textbf{SENTENCE}}$ [$^{\textbf{NOUN 1, Agent}}$ *the dog*] [$^{\textbf{VERB}}$ *chased*] [$^{\textbf{NOUN 2, Patient}}$ *the cat*] *up* [$^{\textbf{NOUN 3, Locus}}$ *the tree*] [$^{\textbf{NOUN 4, Time}}$ *this morning*]]. From a forensic linguistic point of view it is important to note that recipients (listeners or readers) inferentially assign semantic roles to noun phrases in texts, and that cut-and-paste plagiarism could be detected by looking for discontinuities between sections of a text where the semantic role for a particular noun phrase is unexpectedly switched with another role.

Humans experience a great variety of interactions between entities in real life, but only use 5 basic patterns to express our thoughts about such events:

Agent dominates Patient

a. The boy broke the plate
b. The lady is wearing a blue dress
c. John wrote a poem
d. Sally sang a song

Co-agent cooperates with co-agent

a. John and Peter are taking (with/to one another) *or* John is talking to Peter
b. John and Peter are helping one another or John is helping Peter

Counter-agent competes with counter-agent

a. Sue and Jane are competing/arguing/debating (with one another) or Sue is competing/ arguing/debating with Jane
b. The dog and the cat are fighting (with one another) *or* The dog is fighting with the cat

Experiencer experiences stimulus

a. Jack loves Judy
b. Andy hates peas

Stimulus stimulates Experiencer

a. Peas nauseate Andy
b. Horror movies frighten young children

The being, doing and happening schemas present relatively mundane, background scenarios, used to set the scene for the more interesting interactions that involve agents and patients co-agents and counteragents, experiencers and stimuli. We use the being schema to simply situate entities in time and space. We use the happen schema to portray passive processes. We use the do schema to portray individuals engaged in activities on their own. By contrast, we use the different types of agency schemas ([sole] agent, co-agent and counteragent) to portray human interactions, and the stimulus-and-experiencer schema to portray the psychological effects external stimuli on experiencers' mind states. The schema that interrelates stimulus and experiencer is fundamental to cognition and to the interpretation of the narratives that we tell one another. Whenever we try to make sense of what we observe around us, or interpret what others are communicating to us, we are experiencers, subject to stimuli that influence our perceptions and conceptions through our senses.

Humans have an anthropocentric perspective of our environment. We mostly take for granted the time and place of events, and the instruments that we use, and often leave them out of sentences—the most basic mini-stories that we tell one another. We populate our sentences, and the narratives that we weave by combining sentences, with types of agents and patients that we construe as heroes, villains and victims. In sentences agents, patients, stimuli and experiencers are obligatory roles, while it is optional to stipulate instruments, time and place, as in *He sliced the cake (in the kitchen) (with a knife).* Instruments can be foregrounded by using them in theme position at the head of sentences, as in *the dog fetches the paper <u>every morning</u>*, which becomes *<u>every morning</u> the dog fetches the paper*. Similarly, instruments can be foregrounded by using them in the theme position, as in *he killed his opponent <u>with this dagger</u>* which becomes *<u>with this dagger</u> he*

killed his opponent. Instruments can also be foregrounded by reconceptualising them as agents, as in *this dagger killed his opponent.* When an instrument is promoted to the role of the agent, it is used in the subject position at the head of the transitive sentence.

Plagiarised texts often have a feel of ill fitting sentences and paragraphs because the semantic role relationships that were used to deploy the theme of the original text do not fit together in the plagiarised text.

Text Cohesion: Phoric Reference Relationships within and between Sentences

Phoric referencing relates to the use of provisional text elements (pro-forms) like pronouns and pro-adverbs that are used in texts to set up thematic reference chains throughout texts by referring back to previous text elements or by referring forward to subsequent text elements. Halliday & Hassan (1976), Van Hoek (1992; 1997), Liddy *et al.*, (2007), Lyons (1977), Beaver (2003), and Wikipedia (http://en.wikipedia.org/wiki/Pro-form) present good introductions to phoric text referral. The most commonly used form of text reference is anaphora, or backward referencing. Table 2 provides examples of phoric text reference in English:

Table 2: Anaphoric, cataphoric and exophoric text referencing by pronouns

Pronoun	**Explanation**	**Example**
Anaphora	Pronoun refers back to prior noun/ noun phrase/ clause	'This' in: We're lost. *This* is our problem. 'he' in: John said *he* will come
Cataphora	Noun refers forward to subsequent noun/ noun phrase/ clause	'This' in: *This* is our problem. We're lost. 'he' in: *He* will come, John said.

Exophora	Noun refers to item/s outside text	'This' in: Take *this* to your mom. 'they all' in: *They all* ran away.

Table 3: Anaphoric, cataphoric and exophoric text referencing by pro-adverbs

Pro-adverb	**Explanation**	**Example**
Anaphora	Pro-adverb refers forward to a prior adverbial element	'Afterwards in: Now sleep. *Afterwards* rest. 'When' in: Immediately go and call me *when* you have arrived. 'subsequently' in: He resigned and *subsequently* emigrated to Australia.
Cataphora	Pro-adverb refers to subsequent adverbial element	'When' in: *When* you are home safely, call me
Exophora	Pro-adverb refers to item/s outside text	'When' in *When we were young…*

Less commonly used than anaphora is cataphora, where a provisional form refers forward to a subsequent text element as shown in Table 3 above. Plagiarism that is the product of cutting and pasting from a variety of documents often is signalled by improper phoric reference links.

Other language forms of cataphoric reference that provide text cohesion, and that are often disturbed during plagiarism:

- **Enumeration** as in: firstly… secondly… thirdly…, a)… b)… c)…, 1.1… 1.2… 1.3…

- **Previewing** as in: 'We will first define the key concepts of the study before proceeding to discuss them', 'The following six symptoms could occur during dehydration:… '
- **Scope setting** as in 'In general, … More specifically…'
- **Reinforcing** as in: 'In summary…', 'In conclusion…' , 'when all is said and done…'

Using Frequency of Usage as a Predictor of Likely Authorship

We have seen that for written language the term 'grammatical pattern' relates to the combinatory principles that are used to generate lexemes (e.g. *cat* and *uneducated*), phrases (e.g. noun phrases like *the dog* and *the tall man on the horse* or prepositional phrases like *in the air* and *up against the fence*), and clauses (e.g. *we rested* or *pick up the tin can that you tossed*). Lexemes essentially are a combination of written signs (letters, numbers, diacritic signs) that form symbols (l+y+n+x forms the word label 'lynx'), which is then associated with clusters of concepts as shown below:

> **Lynx: '**The lynx is a medium-sized cat characterized by its long ear tufts and short (bobbed) tail with a black tip. It has unusually large paws that act as snow shoes in very deep snow and its thick fur and long legs make it appear larger than it really is' (Defenders of Wildlife (2009).

The above definition of the word 'lynx' shows that categorization is at the heart of lexicography. It firstly reveals that we are dealing with a subcategory of cats, a particular type of cat, and secondly that we distinguish this type of cat from other types of wild cats like bobcats, caracals, cheetahs, cougars, genets, leopards, lions, lynxes, ocelots, panthers, and tigers, based on specific distinguishing attributes like ear tufts, a shot bobbed tail and large paws.

An important aspect of lexicography is that lexemes that belong the same subcategory, as in [**Cat** [**House Cat** …] [**Wild Cat** bobcat, caracal, cheetah, cougar, genet, leopard, lion, lynx, ocelot, panther, tiger …]], could be used at different frequencies by everyday language users as can be seen in Table 4:

Table 4: Frequencies with which lexemes referring to different types of wild cat are used on the Internet, based on a Google search

Types of Wild Cat	**Google Hits (Million)**	**Types of Wild Cat**	**Google Hits (Million)**
caracal	1,69	genet	9,88
ocelot	2,21	cougar	13,8
bobcat	5,28	panther	22,1
lynx	6,1	leopard	40,2
lion	7,6	tiger	138
cheetah	9,48	genet	9,88

Lexemes with a low frequency of usage tend to be better predictors of likely authorship than ones with a high frequency of usage when used in phrases as can be seen from the Google search results for the phrases *land of the tiger* and *land of the caracal* as seen in Table 5:

Table 5: Frequencies with which the lexemes 'tiger' and 'caracal' are used on the Internet in the phrase 'land of the X', based on a Google search

Phrase	**Google Hits**
'Land of the tiger'	61,600
'Land of the caracal'	0

Given that two lexemes can share the same subcategory in the English lexicon, the knowledge domain in which a lexeme is used, it is important for forensic audit purposes to establish the frequency of usage of lexemes. For instance, in the domain of narrative fiction two lexemes are used to refer to a person who invents or recounts fables, namely *fabler* and *fabulist*. While these two synonyms are semantically equally weighted in the English lexicon, *fabler* is used with a significantly higher frequency than *fabulist* as can be seen in Table 6:

Table 6: Frequencies with which the lexemes 'fabler' and 'fabulist' are used on the Internet in the clause 'he is a X', based on a Google search

Phrase	**Google Hits**
'he is a fabler'	165,000
'he is a fabulist'	2

Consider the following forensic linguistics auditing scenario: the lexeme *fabulist,* which has a low frequency of usage in texts, is encountered in the same syntactic context in two different manuscripts of which one author alleges plagiarism and the other author alleges independent authorship. If a third manuscript, predating the two in question is encountered in which the accused author is also using the term *fabulist* a good case could be made for independent authorship. If, however, the third manuscript has been written by the author who alleges that s/he has been plagiarised, a good case would be made to support the plagiarism allegation.

Nonstandard Language Forms as Predictors of Likely Authorship

When considering the utility of grammatical patterns to attribute likely authorship of documents a clear distinction has to be drawn between different *varieties of written language*. One can for instance distinguish the standardised variety of a language that is used in formal settings like when one writes an academic paper, a thesis or a funding proposal. In such formal settings full forms of words, like *cannot, will not* and *is not* tend to be used, rather than contracted forms like *can't, won't* and *isn't.*

There however, also are a number of nonstandardised varieties of a language like regional dialects and social dialects that must be kept in mind during forensic auditing, where nonstandard variants like *cain't* and *ain't* appear in writing, or the nonstandard dialect variant *brung* for the past tense of the verb *bring* rather than the standardised variant *brought*. Such nonstandard varieties of language are considered to be markers that may reveal the social dialect, regional dialect or the idiolectic (individualistic) affiliation of a language user.

It is also important to note that during *spontanious communication* in informal settings, like when people are excited, angry or emotionally upset they tend to use language patterns *subconsciously*. By contrast, in formal settings, like when one is reading a written a speech, while one is answering questions during a an interview, or during a written examination one tends to *self-monitor* one's language output, often tending to reformulate or further clarify earlier statements.

From the above examples it should be clear that individual language users may subconsciously leak language forms that will could be used by forensic investigators to attribute likelihood of authorship of written documents.

The Forensic Audit Process

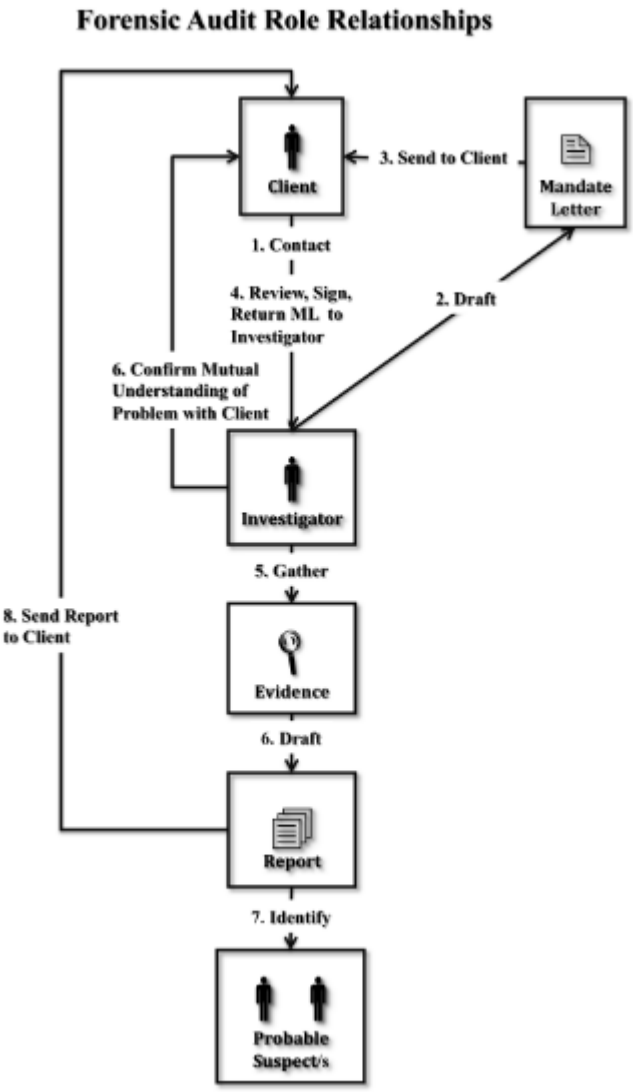
This section outlines the forensic audit process according to which shallow and deep web search protocols will be implemented to retrieve and analyse possible source texts that have been plagiarised.

The first aspect of such a forensic audit that has to be noted is that the process requires the systematic implementation of a systematic pre-planned procedure.

The second aspect is that each phase of the audit should be meticulously noted because such notes form the basis of professional reports, and assist the auditor to coherently answer questions in case of subsequent legal proceedings.

The third aspect that is important is that due to the transactional nature of forensic audits, semantic role analysis works well to characterise the forensic audit process itself. The same semantic roles that were identified as part of text analysis apply to the forensic auditing process, because the auditor acts as an agent on behalf of a client after having received a mandate from the client to investigate the likely authorship of a document, using a variety search instruments to locate relevant documents on the shallow web as well as the deep web, in order to determine whether a suspect legally is the actual author of a document, or merely a plagiarist. The role-based forensic audit procedure is shown in Figure 9:

Figure 9: The participant role relationships in forensic auditing



Conclusions

In this contribution plagiarism has been identified as one of the major problems that plague present-day academia. A problem-solution oriented approach was implemented to identify the appropriate electronic resources that could be used to retrieve potential electronic source documents for forensic audits to establish likelihood of authorship. The information reviewed enabled me to answer the three research questions posed at the beginning of the article, namely:

1. Where on the Internet do plagiarists obtain plagiarised sources?
2. What resources can be used to track and eliminate plagiarism?
3. What methods of text analysis could serve as forensic tools to identify plagiarism on linguistic grounds?

With regard to research question 1 (relating to the sources of plagiarised material) it was shown that plagiarists can obtain sources from the shallow web as well as from the deep web.

With regard to research question 2 (relating to the resources available for tracking and eliminating plagiarism) the ineffectiveness of current algorithm-based anti plagiarism tools was highlighted, additional computer-based as well as online resources and services were identified.

With regard to research question 3 (relating to the methods of text analysis that could be used as forensic audit tools) it was emphasised that a nuanced analysis of documents requires semantic-based analytical procedures to help distinguish between authors and plagiarists. An overview was provided of the lexical, semantic, syntactic, variational and stylistic aspects of language that would enable auditors to make a nuanced semantic analysis if documents, also with regard to likelihood of authorship.

Finally, the case was presented for combining Cyber Forensics and Forensic Linguistics into a coherent interdiscipline that would enable forensic auditing of documents on semantic grounds. Such a course of studies would ensure that relevant knowledge in informatics, linguistics and auditing can be harnessed to counteract plagiarism and promote ethical academic writing.

References

Academic Index: http://www.academicindex.net.

Animal Omnibus No Date. *Mammals.* Accessed online on October 2009 at http://www.animalomnibus.com/.

Article Checker 2008. Article Checker. Available at http://www.articlechecker.com/. Accessed 28 October 2009 13:53.

Baldwin Timothy, Valia Kordoni and Aline Villavicncio 2009. Prepositions in Applications: A Survey and Introduction to the Special Issue. *Computational Linguistics* 35,2: 120-149. Association for Computational Linguistics.

Baldwin Timothy, John Beavers, Emily M. Bender, Dan Flickinger, Ara Kim and Stephan Oepen 2009Beauty and the Beast: What running a broad-coverage precision grammar over the BNC taught us about the grammar—and the corpus. University of Washington Homepage. Accessed online on 15 October 2009 at http://faculty.washington.edu/ebender/papers/beauty_beast_final.pdf.

Beaver, David I 2003. The Optimization of Discourse Anaphora Presente at the Linguistic Society of America Annual Meeting (San Francisco, 2002). ©2003 Kluwer Academic Publishers.

Bergman, Michael K 2009. The Deep Web: Surfacing Hidden Value. Bright Planet Corporation. Original publication date, July 2000; minor revisions 2009.

Carrier, B & E Spafford 2004. An Event-Based Digital Forensic Investigation Framework. Presented at the *Digital Forensic Research Workshop (DFRWS).*

Clusty meta-search engine: http://clusty.com.

CopyCatchGold 2009. Available at http://www.copycatchgold.com/. Accessed 3 October 2009 09:21 AM.

CopyScape 2009. Available at http://www.copyscape.com/. Accessed 28 October 2009 20:42.

DBpedia semantic search engine: http://dbpedia.org/About.

Defenders of Wildlife 2009. Lynx. Accessed online on 21 September 2009 at http://www.defenders.org/wildlife_and_habitat/wildlife/lynx.php.

Donneley, Anne 2005. Introduction to Medicolegal Terms. *Journal of the American Association for Medical Transcription* 24,3: 173-175. Accessed online on 21 September 2009 at:

http://www.Jaamtonline.com/pt/re/jaamt/fulltext.01179370-2005060000-0010,htmU.

Dupli Checker 2009. Available at http://www.duplichecker.com/. [Accessed on 27 October 2009 18:57].

Eve2 Essay verification Engine [No Date]. Available at http://www.canexus.com/. Accessed on 27 October 2009 22:00].

Evri semantic search engine: http://www.evri.com.

Forensic Magazine 2009. Vicon Publishing Inc. Accessible at http://www.forensicmag.com/digital.asp.

Freebase semantic search engine: http://www.freebase.com.

Glatt Plagiarism Screening Program [No Date]. Available at http://www.plagiarism.com/screen.id.htm. Accessed on 16 October 2009 18:21.

Gnod semantic search engine: http://www.gnod.net.

Haas, Lura [no date]. Beauty and the Beast: The Theory and Practice of

Information Integration. IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120.

Hakia semantic search engine: http://www.hakia.com.

Halliday, M and R Hasan 1976. *Cohesion in English*. Harlow: Longman.

Hattendorf, John B (editor in chief) 2005. Description of the Project. *Encyclopedia of Maritime History*. Oxford University Press, USA. Accessed online on 21 September 2009 at http://www.us.oup.com/us/brochure/0195130758/.

Holcombe, C. John 2007. Linguistics. Textetc.com. ©Litlangs Ltda. 2004 2005 2006 2007. Accessed online on 18 July 2009 at http://www.textetc.com/theory/linguistics.html.

ISACA 2004. IS Auditing Guideline—Computer Forensics. Document G28. Accessed online on 18 September 2009 at http://www.Isaca.org.

Johnson, Mark 1987. *The body in the mind: The bodily basis of meaning, imagination, and reason.* Chicago: University of Chicago Press.

Jplag [No date]. *University of Karlsruhe*. Available at https://www.ipd.uni-karlsruhe.de/jplag/. Accessed on 12 October 2009.

Kartoo semantic search engine: http://www.kartoo.com/flash04.php3

Kersaint, G, B Horton, H Stohl & J Garofalo 2003. Technology Beliefs and Practices of Mathematics Education Faculty. *Journal of Technology and Teacher Education* 11,4: 549-577.

Klopper, Rembrandt 1999. How Science Reveals The Universe, And How Humans Perceive The Universe To Be. *Alternation* 6,2: 248-272.

Klopper, Rembrandt 2003. Yesterday is Another Country …Image schemas in conceptual blending to optimize human scale thinking. *Alternation* 10,2: 1023-1757.

Klopper, Rembrandt 2008. How to Organize Your Dissertation. Research Training Document. School of Information Systems & Technology, University of KwaZulu-Natal.

Lamb, Sydney M 1998. *Pathways of the Brain: The Neurocognitive Basis of Language.* Amsterdam: John Benjamins Publishing Company.

Liddy Elizabeth, Susan Bonzi, Jeffrey Katzer and Elizabeth Oddy 2007. A Study of Discourse Anaphora in Scientific Abstracts. *Journal of the American Society for Information Science and Technology* 38,4:255-61.

Lyons, John 1977. *Semantics.* Volumes I, II. Cambridge, England: Cambridge University Press.

Mandler Jean M 2004. Thought before Language. *TRENDS in Cognitive Sciences* 8,11:508-513

Marcella, A & G Greenfield Robert 2006. Cyber Forensics. Accessed online on 15 September 2008 at http://www.cyber-forensic-analysis.com/.

Markman, Ellen M 1990. Constraints Children Place on Word Meanings. *Cognitive Science* 14:57-77.

McMenamin, G, Dongdoo Choi, PJ Mistry, Susan Morton & Wakako Yasuda 2002. *Forensic Linguistics: Advances in Forensic Stylistics.* Boca Rotan, London, New York: CRC Press.

Miller, Alisa [no date]. 100 Useful Tips and Tools to Research the Deep Web. Online College Blog. Online College Reviews. Accessed online on 15 October 2009 at http://www.online-college-blog.com/index.php/features/100-useful-tips-and-tools-to-research-the-deep-web/.

Moral Philosophy.Info. [No Date]. Accessed online on 14 December 2009 at http://www.moralphilosophy.info/.

Pearson Education 2004. Conceptual Framework. *Glossary.* Accessed online on 21 September 2009 at http://wps.pearsoned.co.uk/wps/media/objects/1065/1090612/glossary.html.

Pl@giarism 2008. Accessed 29 October 2009 19:20 at http://people.few.eur.nl/span/Plagiarism/index.htm.

PlagiarismDetect 2009. Accessed 25 October 2009 15:56 at http://www.plagiarismdetect.com/.
Powerset semantic search tool: http://www.powerset.com.
Quintura semantic search tool: http://www.quintura.com.
Quirk Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik 1974. *A Grammar of Contemporary English*. London: Longman.
SafeAssign 2007. Available at http://www.mydropbox.com/. Accessed 20 October 2009 17:00.
Sampson, Geoffrey 1980. *Schools of Linguistics: Competition and Evolution*. Stanford: Stanford University Press.
Taylor, John R. 2003. *Linguistic Categorisation.* Oxford: Textbooks in Linguistics.
SurfWax: http://www.surfwax.com.
Turbo 10 meta-search engine http://turbo10.com.
Ungerer, Friedrich and Hans-Jörg Schmid 2006. *An Introduction to Cognitive Linguistics.* Harlow/ London: Pearson Longman.
Van Hoek, Karen 1992. *Paths through conceptual structure: Constraints on pronominal anaphora* Unpublished PhD dissertation, University of California, San Diego.
Van Hoek, Karen 1997. *Anaphora and Conceptual Structure*. Chicago: The University of Chicago Press.
Wikipedia 2009. The Free Encyclopedia. Available at: http://en.wikipedia.org/wiki/ Main_Page.
WordNet 2009. Word Search, © 2009 The Trustees of Princeton University. Accessed online on 21 September 2009 at http://wordnet.princeton.edu/.
World Curry Guide: http://web.curryguide.com**.**
Wright, Alex 2009. Exploring a 'Deep Web' That Google Can't Grasp. New York Times. Accessed online on 15 September 2009 at http://www.nytimes.com/2009/02/23/technology/internet/23search.html ?_r=1&ref=business
Zotero semantic search tool: http://www.zotero.org.

Rembrandt Klopper
School of Information Systems & Technology
University of KwaZulu-Natal
Durban, South Africa
rklopper@gmail.com