INVESTIGATING THE USE OF FORENSIC STYLISTIC AND STYLOMETRIC TECHNIQUES IN THE ANALYSES OF AUTHORSHIP ON A PUBLICLY ACCESSIBLE SOCIAL NETWORKING SITE (FACEBOOK)

by

COLIN SIMON MICHELL

submitted in accordance with the requirements
for the degree of

MASTER OF ARTS

in the subject

LINGUISTICS

at the

UNIVERSITY OF SOUTH AFRICA

SUPERVISOR: PROF EH HUBBARD

JULY 2013

Declaration

I declare that “Investigating the use of forensic stylistic and stylometric techniques in the analysis of authorship on a publicly accessible social networking site (Facebook)” is my own work and that all sources that I have used or quoted have been indicated and acknowledged by means of complete references.

Signature:

Colin Simon Michell

[Student No: 4322-607-8]

Summary

This research study examines the forensic application of a selection of stylistic and stylometric techniques in a simulated authorship attribution case involving texts on the social networking site, Facebook. Eight participants each submitted 2,000 words of self-authored text from their personal Facebook messages, and one of them submitted an extra 2,000 words to act as the 'disputed text'. The texts were analysed in terms of the first 1,000 words received and then at the 2,000-word level to determine what effect text length has on the effectiveness of the chosen style markers (keywords, function words, most frequently occurring words, punctuation, use of digitally mediated communication features and spelling). It was found that despite accurately identifying the author of the disputed text at the 1,000-word level, the results were not entirely conclusive but at the 2,000-word level the results were more promising, with certain style markers being particularly effective.

Key words

Facebook; authorship attribution; style markers; idiolect; forensic linguistics; forensic stylistics; stylometrics; WordSmith Tools

Acknowledgements

I would like to express my thanks to the following:

- All the researchers and experts whose work I consulted.

- Professor Hilton Hubbard for being such a supportive and tolerant supervisor, who occasionally allowed me to trip over my own feet until I understood what had to be done.

- The participants who graciously sent me writings from their Facebook inboxes and allowed me a privileged look into their private worlds.

- My wife Nicola and daughter Roshyin and son Kiyan for all their moral support.

- My previous employers at International House Johannesburg who gave me time during my work day to visit Pretoria for discussions with Prof. Hubbard, and to my current employers at the Higher Colleges of Technology - Fujairah for allowing me the use of their facilities and photocopiers.

Finally, this research is dedicated to my late father John Howard Michell, who always supported my dreams to better myself, and Roshyin's twin sister Skylar, who was with us for such a short time.

Table of Contents

List of figures

List of Tables

Chapter 2 – Literature review

Chapter 3 – Research method

Chapter 4 – Findings

Chapter 1 – Introduction

> *How reliably can linguistic experts establish that Person A wrote Document X when document X is an e-mail – or worse, a terse note sent by instant message or Twitter? After all, e-mails and their ilk give us much more limited purchase on an author's idiosyncrasies than an extended work of literature. Does digital writing leave fingerprints?*
>
> (Zimmer 2011, 12)

1.1 Introduction

Within the last 10 years social networking sites such as Facebook, MySpace, Twitter and LinkedIn have exploded into the public domain and have become a multi-billion dollar industry worldwide courted by governments looking for re-election, Arab revolutionaries seeking regime change, disaffected youth in London inciting riots and ordinary people reaching out to friends and family. For linguists, social networking sites and electronic communication in general have brought new forms of language use, and for forensic linguists the question has arisen as to how feasible it is to attribute authorship to a single author within a group of possible authors in this new form of communication.

This chapter begins by introducing the research problem and the research aims, before moving on to the rationale for the study. The rationale begins with an examination of the history of social networking sites, with an emphasis on Facebook, and then moves on to exploring criminal activity conducted with the aid of Facebook, such as children being groomed by paedophiles, identity theft, murder, rape, kidnapping and cyberbullying. Then there is an introduction to linguistic features common to Facebook, and to socially orientated digitally mediated communication more generally, which highlights the need for more forensic linguistic investigation of authorship on social networking sites. Introductory remarks are also made about the research method, which is followed by an overview of this dissertation's chapters.

1.2 Research problem

To what extent is it possible to apply researched forensic stylistic and stylometric analyses in a publicly accessible social networking site such as Facebook in order to reveal sufficient difference in each author's output to assist in attributing authorship? There are two models for the study of variation in language: the bottom-up and the top-down models (McMenamin 2002). The bottom-up model involves looking for recurrent patterns, distributions and forms of organisation so as to find evidence of the existence of patterns and examples of rules relating to the writer's style. The top-down model of stylometrics looks for a "predetermined taxonomy of stylistic items which would allow for the discrimination of writers within a certain community" (McMenamin 2002, 54). In the context of this study, *stylistic* refers to the qualitative analysis, where "linguistic features are identified and then described as being characteristic of an author" (McMenamin 2002, 76). On the other hand, s*tylometric* refers to quantitative analysis, where "certain indicators are identified and then measured", for example by counting the relative frequency at which a feature occurs in a text (McMenamin 2002, 76). In a stylistic approach, the researcher generally employs a bottom-up approach, as he or she needs to analyse the text (usually, but not necessarily manually) for features which are idiosyncratic to that author (see Table 3.1), whereas, in a stylometric approach, the researcher will employ a top-down approach as the identified features are usually from a predetermined list: for example, specific function words or punctuation points. However, it is not a case of using either stylistic or stylometric methods exclusively, but rather in conjunction:

> Qualitative and quantitative methods complement one another and are often used together to identify, describe and measure the presence or absence of style-markers in questioned and known writings. (McMenamin 2002, 76)

Kotzé (2010, 187) asserts that a more holistic approach is needed due to the "multifarious nature of evidence contained in written texts", and the use of both qualitative and quantitative methods in conjunction will result in more reliable authorship attribution. That is why my study, too, uses a combination of both approaches. In the

stylistic examination of the texts I looked for patterns and idiosyncrasies regarding punctuation, typography, spelling, lexis and features associated with digitally mediated communication. The style markers analysed stylometrically are keywords, function words, most frequently occurring words and punctuation. Punctuation is examined both stylistically and stylometrically, due to the nature of Facebook discourse. Firstly, it is looked at stylistically, as many of my participants exhibited idiosyncratic punctuation usage, for example, multiple question marks, and secondly, it is analysed stylometrically, where all punctuation features are counted and statistically assessed.

Having discussed two of the key terms in my dissertation title above, a little more should be said here about a third term with which both are linked, namely 'forensic'. My study is a forensic linguistic one in that it is concerned primarily with examining how similar the texts of different writers are to a hypothetical 'disputed text' that could, in principle, have been the subject of court proceedings, as opposed to just being concerned with stylistic and stylometric descriptions of texts undertaken for other reasons.

My study is concerned with the analysis, for forensic purposes, of the language used in social networking communications, which, like other forms of digital communication used for social purposes, tends to come in short, relatively spontaneous and unedited text, and is closer in style to spoken discourse than to written discourse (Crystal 2011). Despite focussing on Facebook, I will be making reference to other social networking sites such as Twitter and MySpace, as well as other forms of digitally mediated communication such as Internet Relay Chat (IRC), Short Message Send (SMS) and e-mail. Despite very little research having been done exclusively on the language used on Facebook, Crystal (2007) suggests that the register of the writing used in digitally mediated communications for social purposes is sufficiently similar across electronic mediums (e-mail, SMS, IRC, and social networking sites) for them to be considered a single genre. However, in later work Crystal (2011, 77) re-evaluates his initial view by asking: "Is social networking [...] demonstrating 'sub-varieties' or varieties in their own right?". Crystal's (2011) point highlights the evolving nature of communication conducted via electronic mediums, and gives the need for studies focussing on the

language used on Facebook, as well as other social networking sites and excluding electronic communications used for non-social purposes. One needs to remain cognisant of the fact that different contexts influence language features and linguistic expectations, and as Crystal (2007, 7) asserts, “all language-using situations present us with constraints which we must be aware of and must obey if our contribution is to be judged acceptable”.

In sum, then, my research problem concerns the extent to which it is possible to attribute authorship on the social networking site, Facebook, using a combination of stylistic and stylometric methods.

1.3 Research aims

Traditionally, stylistic and stylometric research has been focussed on more formalised types of writing such as letters, books and articles, especially the books of the Bible and Shakespeare’s plays (Morton 1978). Research of this sort that has been done on what could be described as informal types of electronic-based writing, such as e-mail, text messaging, blogging, IM (Instant Messaging) and chat room discourse is somewhat limited and has mainly focussed on text messaging (Olsson 2009), e-mail (Chaski 2005) and MSN chat (Grant 2010).

The overall aim of my research can be broadly stated as follows:

> To explore to what extent it is possible to attribute authorship on a publicly accessible social networking site, such as Facebook to a single author among a group of authors from similar demographic backgrounds.

This overall aim is given further definition by the formulation of two specific research aims below (together with their objectives, which indicate how each aim is operationalised in practice):

Aim 1

To explore the extent to which each member of a set of style markers can effectively identify the writer of the disputed text.

Objective 1(a)

To explore the texts stylistically by looking for patterns and idiosyncrasies in punctuation, typography, spelling, lexis and features associated with digitally mediated communication in the disputed texts.

Objective 1(b)

To analyse the texts stylometrically in terms of four style markers, namely: keywords, function words, most frequently occurring words and punctuation.

The second specific aim is to determine whether the length of text has any bearing. Both Morton (1978) and Chaski (2010) recommend at least 2000 words of text in order to obtain an acceptable result, particularly if one is conducting a stylometric analysis. However, writing texts in digitally mediated communication rarely reach that level (McMenamin 2002). Hence:

Aim 2

To explore the extent to which the length of the texts available is a factor in how effectively the writer of the disputed text can be identified.

Objective 2

To perform the stylistic and stylometric analyses at both 1000-word and 2000-word levels and to compare the results.

1.4 Rationale for the study

Social networking sites, and Facebook in particular, are a fairly recent phenomenon and are largely unresearched from a forensic linguistic perspective. The underlying rationale of this study is therefore to address aspects of this research gap by investigating

authorship attribution in the context of the relatively unexplored communication medium of Facebook. Facebook was chosen, as it is the largest social networking site, but findings with regard to Facebook could be extrapolated to a large extent to other social networking sites such as Twitter and MySpace, and related forms of electronic communication, such as e-mail written for social rather than professional or commercial purposes, and, instant messaging.

The more detailed rationale that follows consists of three distinct parts. The first looks at why this study is important from a law enforcement perspective and describes a number of examples of criminal acts committed via social networking sites. The second part then provides context by looking at the nature of social networking sites: how they operate and what differentiates them from other forms of digitally mediated communication. It includes a brief history of social networking sites in general, and Facebook in particular. The third section deals with the linguistic conventions used on social networking sites and looks at digitally mediated communication (e-mail, texting, instant messaging) within a broader context, highlighting instances where the discourse used on Facebook is similar to, or different from, other forms of digitally mediated communication. Examples of Facebook discourse are taken from the writings submitted by the research participants.

1.4.1 Introduction to social networking sites

In this section, I will give a brief introduction to the world of social networking sites in order to provide the study with a sociocultural context, and to situate Facebook in the world of social networking. By describing the structures of these sites, it is possible to see how a person with criminal intent could take advantage of the system for their own ends.

Social networking sites are defined by boyd (2007) (danah boyd is a social science researcher who prefers her name to be spelled in lower case) as web-based services that allow individuals to, firstly, construct a public or semi-public profile within a contained system, secondly, create a list of other members with whom they share a

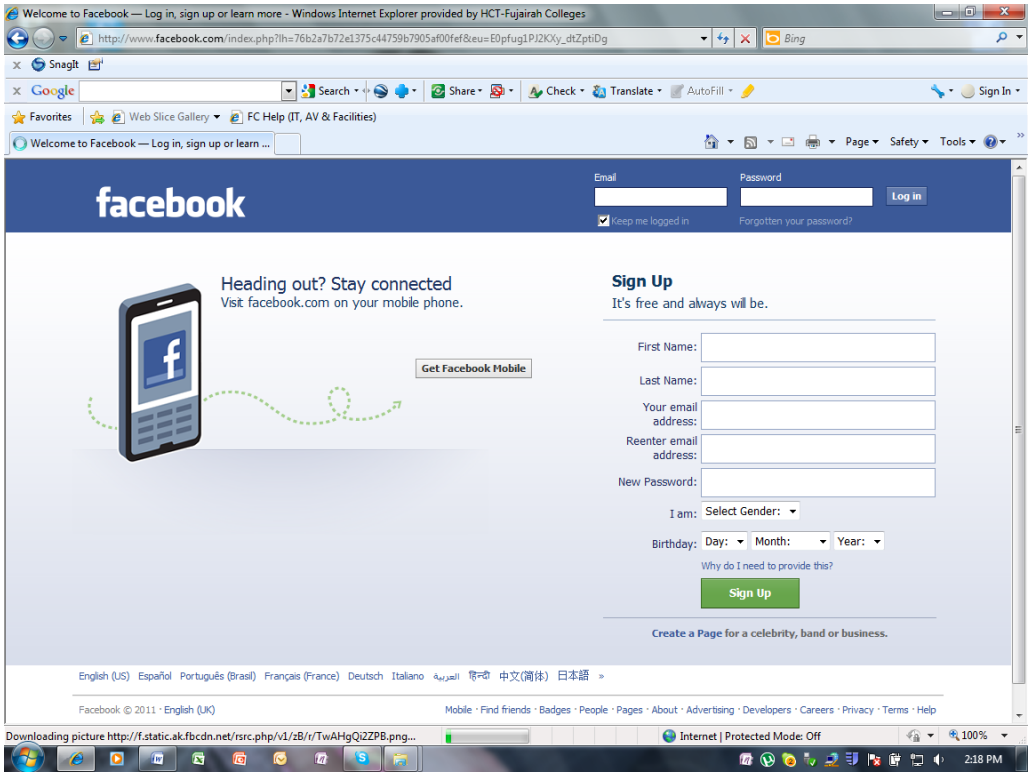
connection, and lastly, peruse their list of connections/friends/followers and those made by others within the system. A unique feature of social networking sites which helps differentiate them from the standard dating sites or school reunion sites is not only that these sites facilitate meetings between people who may have contact with each other, but rather “they enable users to articulate and make visible their social networks. This can result in connections between individuals that would not otherwise be made” (boyd 2007, 1). However, that is not always the aim as these meetings are frequently between ‘latent ties’ (Haythornthwaite 2005) who share some offline connection.

In fact, in most of the larger social networking sites, the members are not looking to meet strangers or new friends or networking in the business sense of the word, but rather to communicate with people who are part of their larger social network (boyd 2007) Facebook is at times the exception to this rule in that once a person joins a group and posts a message on that group’s wall, he or she is likely to interact with people who are complete strangers. This is also common when responding to a person’s status updates. Whereas the initiator of the status update probably knows all the respondents, as they will likely be on his or her friend list, the people responding may not know each other at all. Figure 1.1 is a screen shot of a discussion which resulted from a status update. It is probable that these people have never met and live in different parts of the world. As a result of these interactions, it is possible that they could indeed become friends, or for the criminally minded and paedophiles, this is a gateway to connecting with their intended victim.

Figure 1.1 – A standard discussion post resulting from a status update

Since the inception of social networking sites, starting with Friendster in the early 1990s to the worldwide phenomena that are Facebook, Twitter and LinkedIn, millions of people have integrated these sites into their daily lives and practices (boyd 2007). Acquisiti and Gross (2006) describe social networking sites as sites that allow users to build a profile. Profiles are the unique pages of the individual user where one can “type oneself into being” (Sundén 2003, 5). Profiles will be made public within an enclosed system with the aim of networking with other users of that particular social networking site. This is achieved by allowing users to access information uploaded by other users and permitting other members to access lists of members found on that particular site. Acquisiti and Gross (2006) explain further by stating that the primary goal of social networking sites is to enable people to access pre-existing connections, as well as initiating friendships between people who do not know each other. Generally, all social networking sites operate in a similar manner. New members to a social networking site are required to complete a form, which is followed by requests for information about their personality attributes, preferences, likes and dislikes. Most social networking sites have the option of uploading a personal photograph (boyd and Ellison, 2007). Figure 1.2 is the standard introductory form used by Facebook.

Figure 1.2 – Facebook sign-in page



Even though the key technological features of social networking sites are relatively consistent, it is the cultures that evolve around social networking sites that are the most varied. While most social networking sites help maintain pre-existing social networks, some aid strangers to connect on shared interests such as religion, political views and activities. Some social networking sites appeal to diverse audiences (Facebook, Twitter), whereas others attract people based on mutual interests, common language, shared racial, sexual, religious or nationality based identities (MyChurch, BlackPlanet). (boyd 2007)

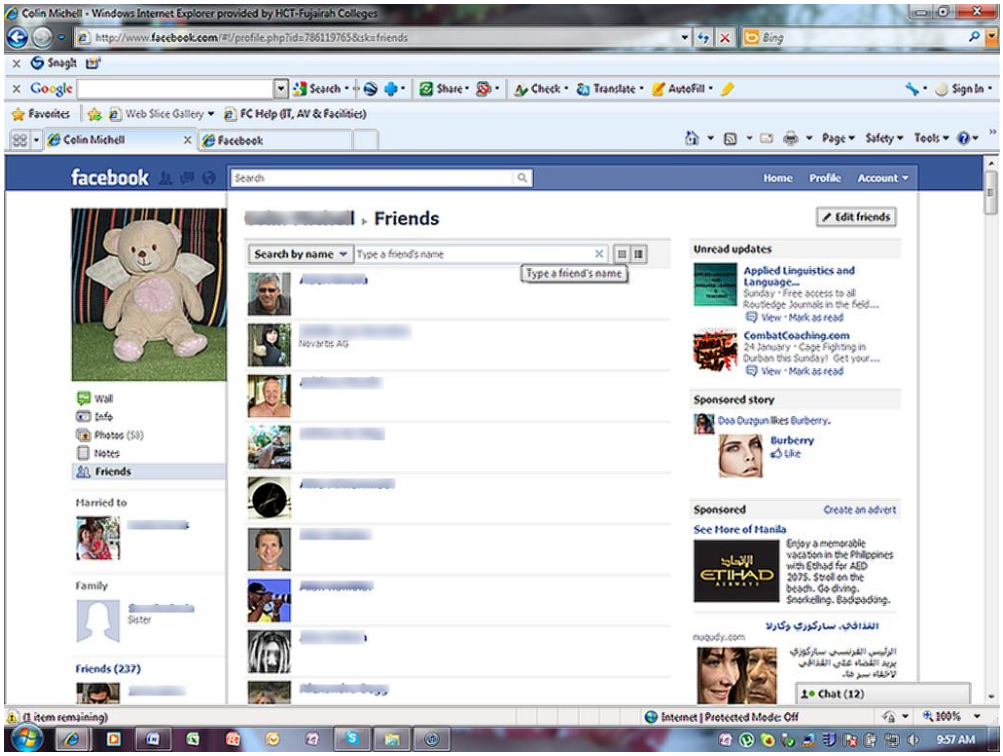
Once a person has joined a social networking site, he or she is encouraged to locate others in the system with whom they have a relationship. The new user can type a person's name or e-mail address into the search engine box to see if the person's friend is in fact on that particular social networking site. Facebook offers the service of trawling

through the new user's e-mail account, looking for names or addresses on the system and offering the chance to send invitations to those not in the system.

However, it is not always possible to view another person's profile without their permission. Both Facebook and Twitter require bi-directional confirmation for friendship (boyd 2007). The label for these relationships differs depending on the site: the most popular include *friend* (Facebook), *follower* (Twitter) and *connection* (Linkedin) (boyd 2007). These labels have even been responsible for adding new words to the English language. The Telegraph (30/12/2009) reports that *unfriend*, meaning to remove someone as a *friend* on a site such as Facebook, was voted the New Oxford American Dictionary's word of the year in America, and is included along with its alternative *defriend* (Savill 2009). In addition to *friend/unfriend*, Facebook has been responsible for other additions to the English language. By 2005, *facebook* had become a verb in its own right to describe the process of perusing other peoples' profiles or updating one's own profile (McDonald 2005) and in 2008, Collins English Dictionary declared *facebook* to be their word of the year (Martine 2012).

One of the most significant aspects of a social networking site is the public display of connections or, in the case of Facebook, the friends list (boyd 2007). This friends list contains links to each friend's profile, which allows a person to "traverse the network graph by clicking through the friend's list" (boyd 2007, 2). On most sites, the friends list is available to anyone who has been accepted as a *friend/follower/connection* to view the profile. In the case of Facebook, unless the privacy settings have been set to restrict non-friends from seeing the friends list, the list is available for anyone to view. Figure 1.3 shows a typical friends list from Facebook.

Figure 1.3 – Example of a friends list



Another prominent feature of social networking sites is the function of leaving messages on friends' profiles. In addition to leaving the message, other members can also leave comments and, in the case of Facebook, there is a *like* facility which can be used by other users who may wish to acknowledge a comment without actually adding anything. A further communication feature of Facebook is that of instant messaging similar to that on webmail and Skype. Facebook also allows users to share photographs and video.

There are over 200 well known social networking sites on the Internet, and since, as of February 2012, Facebook had 845 million subscribers with Twitter trailing behind at 500 million (McNaughton 2012), it therefore seemed judicious to use Facebook as the primary medium for research.

Even though the research in this study is focussed exclusively on Facebook, it is important to situate Facebook within its appropriate context by examining social networking sites in general, and looking at the history of social networking sites before looking at Facebook in particular. Nickson (2009) describes the history of social networking sites. He states that the first true social networking site was SixDegrees.com, which began operating in 1997, and was the first site to allow users to create a profile, invite friends and organise groups, and in 1998 to view other members' profiles.

Despite having attracted millions of users, SixDegrees.com was not able to sustain itself and in 2000 closed down. The founder of SixDegrees.com, Andrew Weinreich, believed that it was simply ahead of its time (boyd and Ellison 2007). This belief was based on the fact that, even though a large number of people were using the Internet, the majority did not have extended networks of friends online, and "early adopters complained that there was little to do after accepting friend requests and most users were not interested in meeting strangers." (boyd and Ellison 2007, 3). The years 1999 to 2001 saw the emergence of community-based social networking sites such as AsianAvenue, BlackPlanet and MiGente. These social networking sites allowed users to create personal, professional and dating profiles. Moreover, users could identify friends on their personal profiles without pre-approval (boyd and Ellison 2007). Other prominent social networking sites to have emerged during this period which allowed users to create visible friend lists were LiveJournal, CyWorld (Korea) and LunarStorm (Sweden). After 2001, a number of social networking sites appeared with the aim of facilitating business connections. The first of these was Ryze, which was aimed at members of San Francisco's business and technology community. This community included the founders of future social networking sites. Andrew Scott, founder of Ryze, describes how the people behind Ryze, Tribe.net, LinkedIn and friendster were all personal and professional friends (boyd and Ellison 2007).

It was only from 2003 that social networking sites became more mainstream and plentiful. The majority of new social networking sites were imitators of Friendster or

attempted to target specific demographics. Essentially, social networking sites could be divided into three distinct groups: (1) socially organised (Friendster), (2) professional (LinkedIn), and (3) 'passion-centric' or niche (Dogster, MyChurch) (boyd and Ellison 2007). In addition to 'pure' social networking sites, websites focussing on media sharing such as Flickr (photographs), Last. FM (music), and YouTube (video) began implementing social networking site features and eventually becoming fully fledged social networking sites (boyd and Ellison 2007).

Facebook was founded by Mark Zuckerberg and three of his college roommates and fellow computer science students, Eduardo Severin, Dustin Moskovitz and Chris Hughes, in February 2004 while attending Harvard as a sophomore. Zuckerberg initially wrote Facemash as the predecessor to Facebook in October 2003, which was a site designed to place photographs of female college students taken from online facebooks, so other students could vote on which one was the most attractive. This prank landed Zuckerberg in trouble with the Harvard administration and Facemash was closed down. However, Zuckerberg gained valuable knowledge and experience from Facemash and on 4 February 2004 launched *thefacebook.* Initially, this site was limited to Harvard University students. Soon after, in 2005, the *the* was dropped from the name. Gradually, other universities were added and in September 2005, anyone with a valid e-mail address and over the age of 13 could sign up (Kirkpatrick, 2010). .

Even though Facebook has many of the generic features found on the majority of social networking sites, such as creating profiles with pictures, listing personal interests, contact information and friend lists, there are a number of features which differentiate Facebook from other social networking sites. The most common is the message system, which is similar to e-mail, where two or more users can communicate privately (see figure 1.4). Secondly, there is the wall, whereby a user can leave a message on another user's wall which is visible to third parties (see figure 1.5). In July 2007, Facebook allowed users to add attachments (hyperlinks, YouTube videos etc.), whereas previously, only text was permitted (Kirkpatrick, 2010). Lastly, there is the

instant messaging (IM) chat service where 2 users can ‘chat’ in real time (see figure 1.6).

Figure 1.4 – An example of Facebook message system

Figure 1.5 – Typical message left on a user’s wall

Figure 1.6 – A typical Instant message thread

1.4.2 Criminal activity on Facebook

The use of social networking sites for criminal activity has in recent years become a contentious issue. Activities range from identity theft, grooming of children by paedophiles (boyd 2007) to the facilitation of rape, kidnapping and murder (Olsson 2010). In addition to criminal activity, Facebook and social networking sites in general have been used as a platform for cyberbullying, which has on a number of occasions had fatal consequences. This section will present examples of how Facebook has been used for the above-mentioned criminal activities. The examples I cite could all have investigative implications for a trained forensic linguist, as proper stylistic and stylometric analysis could aid the authorities in confirming the identity of the perpetrator or add further evidence to the innocence of a falsely accused person.

1.4.2.1 Paedophile activity on Facebook

The Australian parenting website: *raising children.net.au* (2006) reports that the use of on-line chat rooms by paedophiles to groom children is becoming increasingly more widespread. Paedophiles enter into chat rooms pretending to be another child in order to connect with children. These interactions may develop into a relationship and then an

arrangement to meet the child without others knowing about it. Teenagers are especially vulnerable, as they may think they have found a new friend and go to great lengths to keep the relationship private.

The Independent (02/10/2009) reports the case of a paedophile ring where all the members met and shared photographs on Facebook (Taylor, 2009). The MailOnline (28/05/2010) reports how a 28 year old postman posed as a youngster on Facebook in order to groom children aged between 11 and 16 for sex acts (Seamark, 2010). In America, FOXNews.com (06/02/2009) gave a report of an 18 year old male who posed as a female on Facebook and tricked at least 31 male classmates into sending naked pictures of themselves to him, which he then used to blackmail his classmates into performing sex acts with him.

1.4.2.2 Identity theft

MSN BC (2009) reported the case of Bryan Rutberg. Mr Rutberg is a Microsoft employee whose Facebook account was hacked in what is commonly referred to as the Nigerian or 419 scam. Mr Rutberg realised his account had been hacked when his daughter noticed his status update: “BRYAN IS IN URGENT NEED OF HELP”. Within minutes his cell-phone was ringing non-stop with friends offering to help, as many had received e-mails saying that he had been robbed at gunpoint whilst on holiday in the United Kingdom and he needed money to get home. Mr Rutberg then found he had been locked out of his own Facebook account, as the criminals had changed his login details, and had taken the precaution of ‘unfriending’ his wife so he could not post on her wall that he was well and fine.

Figure 1.7 – Part of a chat dialogue between the scammer, a Facebook friend and Mr Rutberg (MSN BC 2009)

Mr Rutberg had been the victim of an updated version of the Nigerian 419 scam, which represents a new trend in computer crime. Instead of sending out millions of spam messages in the hope of trapping one or two gullible victims, computer criminals are becoming more personal by using social networking sites to make their stories more believable. In Mr Rutberg's case, the scammers stole his Facebook identity after discovering his login password, and thereafter changed his page to make it appear that he was in trouble and asked for money. The fact that the pleas for money were made next to his photograph made it even more realistic. Kevin Haley, a director at Symantec Corp's Security Response team, reports that his company has seen a spike in attacks on social networking sites as there has been an increase in the amount of phishing for social networking login details. Mr Rutberg believes that the scammers obtained his login credentials via a phishing e-mail. Had Mr Rutberg set his privacy settings to prevent non-Facebook friends from seeing his e-mail address, he may not have been scammed.

Messmer (2007) reports that users of social networking sites are readily giving up personal information, which could result in them becoming victims of identity theft,

according to a report commissioned by the computer anti-virus software company: Sophos. As an experiment, Sophos created a fictitious Facebook profile called Freddi Stauer (anagram of ID fraudster) and asked 200 random users to be friends with the aim of seeing how much personal information they could collect. Of the 200 requests, 82 accepted, with 72% divulging one or more e-mail address, 84% gave their full date of birth, 87% provided details about education or work, 78% listed their current address, 23% gave their phone number and 26% listed their IM screen name. In addition to that, Sophos gained access to users' photographs of friends and family, as well as personal details such as likes and dislikes and details about employers. One user even disclosed his mother's maiden name, which is a standard security question in most financial and other websites. All of that information was offered freely to someone they did not know and could be used for identity theft.

There is a generally accepted belief promoted by the media that today's youth are not concerned with issues of privacy and will not take steps to protect it (boyd and Hargittai 2010). However, a report entitled *Reputation Management and Social Media,* commissioned by the Pew Internet & American Life Project (Madden and Smith 2010), found that the opposite was in fact the case. Based on research conducted in the autumn of 2009, Pew noted that 71% of 18-29 year old Facebook users surveyed reported that they had changed their privacy settings to increase their privacy levels, whereas only 62% of those Facebook users aged 30-49, and only 55% of those aged 50-64 had done so. It should be noted that the participants of my study fall into the 30-49 age group, which appear to be lukewarm about the need to protect themselves and their privacy online.

1.4.2.3 Murder, rape and kidnapping cases

A recent case in the United Kingdom is that of 33 year old Peter Chapman, who used a fake Facebook profile to ensnare 17 year old Ashleigh Hall from County Durham, whom he subsequently kidnapped, raped and murdered (Carter 2010). Forensic linguist Dr John Olsson was tasked with analysing the mobile message exchange between Ashleigh Hall and Peter Chapman to help reconstruct what happened. The case was

heard at the Teesside Crown Court in March 2010. With the help of Dr Olsson's testimony, a conviction was obtained and Cartwright received a sentence of life imprisonment, of which a minimum of 35 years will be served. In South Africa, a 22 year old man named Thomas Bester (one of his 13 aliases) was sentenced to 50 years in October 2011 in the Durban Magistrates Court for murder, rape, kidnapping and theft. His *modus operandi* was to either meet women in person or befriend them on Facebook under the pretext of being a model scout; once he had groomed them and gained their trust, he would assault them. The most serious of his crimes was the murder of a model in Milnerton, Cape Town, and the rape of two models in Durban. Thereafter, Bester used his Facebook account to taunt the police by saying that they would never catch him.

1.4.2.4 Cyberbullying

A particularly pernicious development in recent times has been that of cyberbullying and cyberstalking. Hinduja and Patchin (2009) define cyberbullying as "when someone repeatedly harasses, mistreats, or makes fun of another person online or while using cell phones or other electronic devices". The Daily Mail (21/08/2009) reported on an 18 year old teenage girl who became the first person to be jailed in the United Kingdom for online bullying after she posted a death threat to a fellow teenage girl on Facebook (Salkeld 2009). The Star newspaper (29/04/2008) covered the rise of online cyberbullying in South Africa, and gave the example of a 16 year old school girl from Randburg who had suffered immense psychological trauma from being bullied online. The same article tells of a 12 year old girl in the United States who committed suicide as a result of being bullied relentlessly online (Beaver 2008). The Cyberbullying Research Center (CRC) in the United States conducted a survey of around 4000 adolescents between the ages of 12 and 18 from a large school district in the southern United States in February 2010. Of the 4000 respondents, almost 20% claimed to have experienced some form of cyberbullying during the 30 day time period. Specific types of cyberbullying include: mean or hurtful comments (13.7%), and malicious rumours spread online (12.9%)(Hinduja and Patchin 2009).

1.4.3 Linguistic conventions on social networking sites

Users of Facebook and social networking sites in general are all part of specific discourse communities, which are defined by Swales (1990, 21) as “communities which have broadly agreed, defined and common goals, special mechanisms for communication and participation and some specialised vocabulary”. Swales (1990) extends this argument by saying that a discourse community is not merely a group of people who share a particular common interest and methods of communicating with each other, but rather one whose participation is determined by that community’s discourse expectations. In order to participate within a discourse community, one must conform to the *norms* (Herring 2007) of that discourse community. Herring (2007) makes use of the term *norms* to describe linguistic and behavioural standards specific to a particular group or discourse community. This section addresses these linguistic issues by examining how online language is evolving due to the rise of digitally mediated communication (DMC). The section starts with an overview of what DMC is, and draws examples of DMC features from the participants of this study’s writings. Thereafter, DMC is examined in a wider forensic context by examining how and why people deliberately obfuscate their identities and how that may affect their language use and the use of cryptolects.

1.4.3.1 What is digitally mediated communication (DMC)?

Crystal (2007) describes the Internet as an electronic, global and interactive medium, and suggests that each of these properties has consequences for the kind of language found there. Crystal (2011) proposed the term digitally mediated communication (DMC), as an alternative to computer mediated communication to describe the language used in an electronic medium as people are now accessing the Internet via mobile phones, iPhones and the like, rather than just through the computer. The language used in DMC for social purposes is often described as ‘written speech’ and users of DMC for social purposes are encouraged to ‘write the way they talk’ (Crystal 2007). Davis and Brewer (1997, 2) describe electronic discourse as “writing that very often reads as if it were being spoken – that is, as if the sender were writing talking”. Crystal (2011, 19), however, states that to think of Internet language as merely ‘written speech’ is far too

simplistic. It would be prudent to have speech and writing as end-points on a continuum and the varieties of Internet language can be “located as being *more or less like speech* and *more or less like writing*”. However, Crystal (2011, 21) does concede that: “On the whole, Internet language is better seen as writing which has pulled some way in the direction of speech rather than speech which has been written down”. However, Crystal (2007) points out that even though spoken language has only a limited presence on the Internet, as technology improves, the use of interactive voice and speech synthesis will give support to the graphic representation.

1.4.3.2 Linguistic implications of digitally mediated communication

Even though the Internet is a relatively recent development, it has come to dominate how we communicate with each other. E-mail, instant messaging and Facebook status updates, amongst others, have added new dimensions to how we communicate with each other. Crystal (2007) has raised the question of how this paradigm shift in communication is affecting language usage. Donath et al (2006, 6) point out that “most on-line conversation is still text”. In this context, text refers to any form of the written word, which is the medium of Facebook, rather than audio or visual media, such as Voice over Internet Protocol (Skype). Herring (2007, 1) extends this notion by saying that “computer [digitally] -mediated communication is predominantly text-based human-human interaction mediated by networked computers or mobile telephony”. While it is true that photo and video have become more common in recent years, the bulk of communication facilitated by computers remains written text. As such, the user is constrained by what can or cannot be done with the keyboard (Becker and Stamp 2005). As with all writing genres, the lack of physical proximity between the interlocutors results in facial expressions, gestures, tone of voice, turn taking and back-channelling devices (*uh uh, yeah*) not being able to be performed whilst the message is being sent (Donath et al. 1999). To overcome these limitations, Internet users have adopted innovative ways to overcome these obstacles by moving beyond conventional ways of writing, by making creative use of punctuation and spellings to show emotions (Walkley 2009). This concept is not entirely new as personal correspondence in the past made use of drawn love hearts, smiley faces and acronyms such as FYEO (*for your eyes*

only). Crystal (2001, 48) suggests that the discourse on the Internet is a “new species of communication” complete with its own lexicon, graphology, grammar and usage conditions. Baron (2003, 88) adds to Crystal’s assertion by saying that “technology often enhances and reflects rather than precipitates linguistic and social change”. In other words, computer mediated communication could be simply mirroring the emerging tendency for written text to become more similar to speech in a process referred to by Leech and Smith (2005) as colloquialisation. Danet (2001) states that multiple punctuation marks, eccentric spelling, overuse of capital letters and asterisks are frequently used to show emphasis, enthusiasm or emotional state. For example, Writer B in my study makes use of multiple punctuations, smiley faces and eccentric spelling all in one sentence.

Example 1

A little birdie told me so sad news!!! is it true you're leaving us????y y y y y y y !!!B sniff sniff sniff....:(

These features help Writer B compensate for the fact that she is not in close physical proximity to her interlocutor and allow her to express paralinguistic features such as tone and emotion (Crystal 2007)

In example 2, Writer H makes use of asterisks to show a physical activity. By placing a word such as *hugs* or *sighs* between asterisks, a person can express a physical activity from the sender, which can also avoid any misunderstanding which may occur due to the inability to see the interlocutor’s paralinguistic signals. Moreover, prosodic features such as stress, tone and loudness can be shown by the use of uppercase (Crystal 2007).

Example 2

So YOU'RE the Patient!! hahaha ...go figure, the psychologist needs his own head checked. :D
hugs *you know I mean well don't you, I really hope it isn't serious.*

It is not unusual for some writers on social networking sites to make use of abbreviations (*c* for *see* and *tnx* for *thanks*) (Crystal 2007). Even though these

abbreviations are more common in mobile phone texting, they are quite common on Facebook and other social networking sites. In example 3, Writer C makes use of eccentric spelling (hue = have) and the abbreviation *gr8* (great), and even *OK* was abbreviated to *K*.

Example 3

K see ya and hue a gr8 w end!

Other features of Facebook writing which are similar to mobile phone texting include the use of lower case when prescriptive grammars and written norms call for upper case, especially with the first person singular pronoun '*I'* which is often rendered as '*i'* (Crystal 2007), as can be seen in example 4 from Writer C.

Example 4

Hey Gav, all is going well thanks :) Though i almost died today! Had a blow out on the highway. 2 guys (i'd call them my angels) stopped to help me - done in 10 minutes then i was back on the road.

Baayen et al. (2000) assert that messages in DMC tend to be quite short, generally speaking, less than 100 words, which may be true for mobile phone texting, instant messaging and status updates on Facebook. Yet messages sent to personal inboxes are frequently much longer than 100 words. I noticed in the texts from my participants that many messages are frequently over the 200 word mark, although a fair number are still under 100 words. However, even if the individual messages are short, this is compensated for by the fact that there is usually a thread of messages as the two parties continue the conversation.

1.4.3.3 Online anonymity and cryptolects

An important aspect of DMC, especially within a forensic linguistic context as discussed by Crystal (2011), is the issue of anonymity. Linguists have always placed emphasis on the situational factors which motivate or inform what language is used. Factors such as age, gender, class and ethnicity are considered crucial pieces of information, yet the Internet is a medium where participants can conceal their identities, and any disclosed information should be treated with suspicion. Even distinctions, such as whether the

user is male or female, or a native or non-native user of a language can be obfuscated. It is true that the Internet is not the first medium of communication to allow anonymity, but it is the first to allow it on such an unprecedented scale and in so many different media, especially chatrooms, blogging and social networking. It stands to reason that this anonymity has an effect on the linguistic output of DMC participants.

> Operating behind a false persona seems to make people less inhibited: they may feel emboldened to talk more and in different ways from their real-world linguistic repertoire. They must also expect to receive messages from others who are likewise less inhibited, and be prepared for negative outcomes. There are obvious inherent risks in talking to someone we do not know, and instances of harassment, insulting or aggressive language, and subterfuge are commonplace. (Crystal 2011, 14)

Apart from the fairly generic features described in the previous examples, there are instances where social networking site users actively utilise a cryptolect, which is defined by Hancock (1986) as a secretive language used by different subcultures, to mask potentially incriminating activities. Lisa Whittaker, a postgraduate student at the University of Sterling, conducted a study on the language used by teenagers aged between 16 and 18 in an urban Scottish district on social networking sites to keep their activities private (The Telegraph 26/04/2010). Whittaker found that instead of admitting to having been intoxicated, they posted *getting MWI* (mad with it). Being *taken* or *ownageeee* indicated that they were in a relationship. Whittaker observed that the language used by these teenagers was not necessarily the same as the abbreviations used for text messaging, whereby all the vowels are removed e.g. txtng (texting), but rather a deliberate attempt to creatively misspell words, thereby strengthening in-group solidarity and concealing their activities from the out-group.

Unfortunately, the use of cryptolects in DMC is not restricted to teenage shenanigans and there have been instances of real criminal intent. Coulthard et al (2011) describe a case where linguists were asked to ‘translate’ the meaning of the following sentence,

which was transcribed from an Internet Relay Chat (IRC) between members of London's Afro-Caribbean community: *ill get da fiend to duppy her den.* The meanings of *da* and *den* are well known, as they are common features of netspeak (Crystal 2007). It was the meanings of *fiend* and *duppy* that required further research. It transpired that *fiend* refers to a drug addict, and *duppy* as a verb means to kill someone in the Jamaican patois. Even though these terms are fairly rare outside of the London Afro-Caribbean community, they are examples of how an in-group cryptolect can be used in DMC for criminal activities.

It could be asked why all this sociolinguistic background is relevant to forensic linguistic and stylometric cases, when so much of the data is analysed quantitatively.

> Sociolinguistic findings about how particular forms are associated with certain social variables can be of great help in forensic linguistic casework, simply because, in addressing authorship issues, the linguist needs to ensure that potentially telling features are individual, i.e. idiolectal, and not dialectal, sociolectal, genderlectal, etc. A linguistic pattern found to occur regularly in the texts under investigation may be genre-specific or the result of accommodation effects. (Coulthard et al. 2011, 536)

The above examples highlight the importance of qualitatively analysing a text in order to find features which are unique or at least peculiar to a specific author. For example, Writer B makes use of the idiosyncratic spelling of *famdamilies* to refer to families. If this were the disputed text, I would be looking for further examples of *famdamilies* in the other texts.

1.4.4 Concluding remarks on the rationale

To conclude the rationale, it can be seen that Facebook, as well as social networking sites in general, have been used for a variety of nefarious activities, ranging from cyberbullying and identity theft to facilitating murder. Another development of DMC has been the advent of new linguistic forms of communication which appear to be

continually evolving as more and more people utilise social networking sites. The meteoric rise of Facebook shows no signs of abating and as of May 2012, Facebook had 900 million subscribers (Sengupta 2012), and it is a fair assumption that criminal activity facilitated by social networking sites will continue, which necessitates the need for continuous research, if law enforcement is to keep pace with the criminal elements who abuse social networking sites for their own ends. This study's focus on applying forensic stylistic and stylometric analyses to Facebook discourse aims to contribute to forensic linguistic research on the effectiveness of such analysis and could ultimately aid in the apprehension of people responsible for criminal activities perpetrated via social networking sites.

1.5 Research method

My research method will be dealt with in detail in Chapter 3, but some initial remarks on it are appropriate here. This study is a simulated forensic case designed to test the forensic linguistic usefulness and validity of a selection of chosen style markers at both the 1000-word and 2000-word level in the Facebook (sub) genre of writing. This research mimics real life forensic linguistic cases, where a disputed text is analysed against a known authored text or texts. To date, I am unaware of any such cases being investigated on social networking sites.

I analysed a 'disputed text' against eight other texts where the authors were known. I obtained 2000 words from eight mother-tongue English females aged between 30 and 40. From one participant I received an extra 2000 words, which represents my disputed text. All the submissions were from their own Facebook messages and were part of their communications to third parties. Only the participants' writings were used and any third party contributions were discarded. The disputed text was then subjected to qualitative stylistic examination of features which appear to be idiosyncratic, and then these were compared to apparently idiosyncratic features of each of the other texts in turn. I then used the concordance program *WordSmith Tools* (WST) to conduct the quantitative stylometric analysis of the texts, firstly for keywords and then to count frequencies of function words, most frequently occurring words, and punctuation marks.

The results were subjected to a Chi-square test, where the profiles of my participants were compared to the disputed text in order to assess the relative effectiveness of the various style markers in correctly identifying the writer of the disputed text.

1.6 Conclusion

With the rise of social networking sites such as Facebook, Twitter and LinkedIn, there has been a corresponding rise of criminal and socially unacceptable behaviour linked to these sites. All indications are that this trend is likely to continue as people become more technologically competent and the technology itself becomes more accessible. Crimes committed via social networking sites typically involve the use of language, and it is this language that can be vitally important evidence in court. Traditionally, stylistics and stylometrics have been concerned with longer texts, such as books of the Bible and Shakespeare's plays, which offer up vast amounts of language for the researcher to analyse. However, in more recent times there has been more emphasis placed on shorter texts such as wills (McMenamin 2002), suicide notes (Shapero 2011) and extortion letters (Hubbard 1995). With the advent of social networking sites, a whole new genre of writing has emerged in response to DMC, which offers the researcher new challenges and opportunities such as new and evolving language conventions. My research is a response to these challenges, as I endeavour to find ways of attributing authorship within a new genre of writing whilst being constrained by fairly small text samples.

1.7 Structure of the study

In concluding this chapter, I will now present an overview of the study by way of chapter outlines.

Chapter 2 examines the literature pertinent to my research. It begins by giving an overview of the history of forensic linguistics, looking at stylistic and stylometric approaches and combinations of the two. This is followed by a review of the linguistic theories which underpin my study, starting with, what has become quite controversial in the forensic linguistic community, the idiolect debate, and then moving on to discuss

language variation and style, as well as differences between spoken and written language. The final section of the literature review discusses research on the style markers that are used in the study, including the stylistic features (punctuation, typography, spelling, lexis and DMC features) and stylometric style markers (keywords, function words, most frequently occurring words, and punctuation). In addition to describing the style markers, the study also reviews some of the contentions surrounding their implementation.

Chapter 3 focuses on the research method. The first section deals with how the participants were selected and the data was collected. It then goes on to describe the various stages that constitute the qualitative analysis, or in other words, the stylistic approach to the study. Thereafter, the study examines the quantitative or stylometric aspects of the study, some of the most relevant workings of *WordSmith Tools* and how it is used to conduct the various tests on the chosen style markers.

Chapter 4 focuses on the findings of my research. The first part looks at the results from the qualitative analyses. Each of the stylistic analyses of the different writers is compared to the disputed text in order to assess how effective different style markers are in identifying which author is the writer of the disputed text. The second main part of the chapter deals with the results of the quantitative, stylometric analyses where again, various style markers are tested for their potential to successfully identify the writer of the disputed text. The first test is one of keyness and involves identifying keywords which distinguish the disputed text from the other texts. Then follow statistical analyses comparing writer X to the other writers in terms of function words, most frequently occurring words, and punctuation. Punctuation is analysed both stylistically and stylometrically. Due largely to the creative aspects of Facebook discourse, non-standard punctuation usage is common and is often used idiosyncratically by an individual writer, thus calling for stylistic examination. Stylometric analyses, on the other hand, can throw light on the overall relative frequencies of punctuation features in the different writers' texts.

Chapter 5 sums up the research with a discussion of its contributions as well as its limitations and looks at possible future research initiatives.

Chapter 2 – Literature Review

> *Everyone knows that language is variable. Two individuals of the same generation and locality, speaking precisely the same dialect and moving in the same social circles, are never absolutely at one in their speech habits.*
>
> (Sapir 1949, 147)

2.1 Introduction

As this research project intends to investigate the use of stylometric analyses to determine authorship on the publicly accessible social networking site Facebook, the literature review will focus on the stylometric and forensic linguistic research which has been done in the past, albeit in different contexts and genres such as novels, letters and e-mail. However, before any stylometric research can be done, it is important to examine the wider sociolinguistic issues of sociolects, idiolects, inter and intra author variation, as well as the unique genre of discourse used on social networking sites, and how it differs from, or is similar to, other digital modes of communication such as e-mail and instant messaging (IM).

This chapter will begin by first giving a brief history of forensic linguistics and stylometry in order to better understand the historical contexts which have shaped our current understanding. This section will begin by describing linguistic disputes regarding the Greek tragedies and then move on to the arguments over Shakespeare's works. Lastly, there will be an examination of the more modern cases of Timothy Evans, Derek Bentley and the Unabomber. The history section will conclude by looking at two notable failures within forensic linguistics, namely the CUSUM method and the authorship attribution of the *funeral elegy*. The second section will examine the field of forensic linguistics and stylometry, by first looking at the nature of writing within a forensic linguistic context. Thereafter, the discussion will move onto the idiolect debate, followed by a discussion on language style and variation. The third section of this review focuses on the practical aspects of style markers, starting with an overview of style markers in general, before looking more deeply into the individual style markers that will be

employed in my study, namely: function words, punctuation and spelling, keyness and commonly occurring words.

The final part of this literature review will focus on the language employed in electronic communication and the nature of digital mediated communication, with particular emphasis placed on how this genre is different to other genres of writing, such as novels and letters. I will also look at how the language employed in social networking sites and electronic communication in general is evolving.

2.2 A brief history of forensic linguistics

Even though the term *forensic linguistics* is a fairly recent development, interest in how language has been used in legal and forensic contexts can be traced back to ancient Greece and Rome (Coulthard et al. 2011). There has been speculation as to whether Homer wrote both the *Iliad* and the *Odyssey*, since both are generally attributed to a single author – Homer, yet both are the result of extensive oral traditions. The Christian Bible has been a focus of linguistic disputes concerning the authorship of all the New Testament letters of St Paul and the Book of Hebrews. Even Shakespeare has come under suspicion with the assertion that Bacon and Marlowe may have contributed to, or completely written a number of his plays. It has even been suggested that *Shakespeare* may have been a *nom-de-plume* for a group of writers (Holmes 1994).

In 1968, a Swedish linguist named Jan Svartvik published *The Evans Statements: A Case for Forensic Linguistics*, wherein he showed that the four statements made by Timothy Evans to the police, regarding the murders of his wife and daughters, “had a grammatical style measurably different from that of uncontested parts of a statement and thus a new area of forensic expertise was born” (Coulthard and Johnson 2007, 5). Timothy Evans was posthumously pardoned 16 years after being executed for murder in 1950 (Coulthard et al. 2011).

A similar case of disputed confession is the Derek Bentley case. Derek Bentley was an illiterate man with a low IQ, who together with another man was involved in an armed

robbery where a policeman was shot and killed. Despite conflicting ballistic evidence and procedural inconsistencies, Bentley was sentenced to death and he was hanged in 1953. Part of the evidence used against him was his confession statement, which had allegedly been transcribed verbatim. Upon reopening the case, however, it was found, for example, that the frequency and usage of the word *then* in the police transcripts showed evidence of 'police language' embedded in the confession, which therefore meant that they were not verbatim transcripts. Bentley was posthumously pardoned in 1998, 46 years after the guilty verdict (Coulthard 2000).

Between 1978 and 1995 Theodore Kaczynski, commonly known as the Unabomber, conducted numerous bombing attacks on universities and airlines. He said he would only cease his bombing campaign if his 35 000 word anti-industrialist manifesto was published in major newspapers. When FBI agents searched Kaczynski's home, they found hundreds of documents authored by Kaczynski which had never been published. When the documents were analysed alongside the manifesto, it was found that there were a number of linguistic features and expressions which appeared in both documents, and despite some features being more distinctive than others, the prosecution put forward the argument that: "the more common words and phrases being used by Kaczynski became distinctive when used in combination with each other" (Coulthard 2000).

Despite the numerous successes of stylometry, there have been a number of failures. Arguably, the most controversial is the CUSUM method, which is an abbreviation for **cu**mulative **sum**, developed by A Q Morton, and was originally used in analysing Biblical texts. Morton based his analysis on the sentence as opposed to the text, in order to calculate the frequency of occurrence of variables such as *number of nouns, words beginning with a vowel, words consisting of three or four letters* and *words consisting of two or three letters*. Morton compared these measurements to the sentence length, which was calculated according to the number of orthographic words (Coulthard and Johnson 2007). Unfortunately, the accuracy of CUSUM was called into question as it was believed that the theoretical framework was not well grounded and

the results were not accurate enough to be considered for criminal matters where peoples' liberty and livelihoods were at stake (Juola 2006 and Stanford, Aked, Moxey and Mullen 1994). The death blow for CUSUM came when Morton was challenged live on British television to attribute authorship to texts he had never seen. Despite his computer program and statistical analyses, it appears to have been an unmitigated disaster, as "Morton could not distinguish between the writings of a convicted felon and the Chief Justice of England" (Grieve 2005, 49). Holmes (1998, 113) went so far as to state that "if stylometry had its 'dark age' then surely this must be it".

2.3 Stylistics and stylometrics within the field of forensic linguistics

McMenamin (2010) describes *stylistics* as the study of style in a language, which he then divides into two sections: literary stylistics and linguistic stylistics. He sees literary stylistics as traditionally concerned with aesthetic and (rather problematically) linguistic conformity issues. Linguistic stylistics, on the other hand, is the analysis of observed style markers as used by groups and individuals. Such stylistic descriptions are often referred to as qualitative analysis.

Burrows (1992) describes *stylometrics* as a development of literary stylistics, which has at its core the assumption that all authors have distinctive writing habits. These writing habits can be exhibited in features such as core vocabulary use, sentence complexity and phraseology, and these features can be categorised and counted. An important assumption is that these features are unconscious habits which are well ingrained. Moreover, stylometrics is concerned with locating textual features which can be used for determining authorship of a text. This is achieved by having a sample of known authored texts from different authors which can be compared to a disputed text. Stylometrics is generally concerned with quantitative analysis.

McMenamin (2002) states that authorship identification is accomplished through the analysis of style in written language, which hinges on the two principles of inherent variability in language: (1) no two writers of a language write in exactly the same way; and (2) no individual writer writes the same way all the time. McMenamin (2002) goes

on to describe the practical applications of studying the underlying linguistic patterns which are used habitually by an author. He suggests that the results of the analysis may be used for: (1) determination of resemblance of questioned writings to a canon of known writings; (2) elimination or identification of one or more suspect authors, and lastly (3) provision of support for neither elimination nor identification.

McMenamin (2002) states that the approach to determining authorship is based on two facts. Firstly author-specific linguistic patterns are present in unique combinations in the style of every writer, and these underlying patterns are usually established enough to be empirically analysed to make identification possible. Secondly, even though a language is *owned* by its entire group of speakers, it is uniquely *used* by individuals in that group. Hubbard (1995, 57) explains that these features are "more like subconscious, automatic habits that develop and become typical of different individuals", much like idiosyncratic paralinguistic features and body language. The reasons why a writer chooses one linguistic form over another is the result of individual preference or habit, and the task to be performed. Therefore, a writer makes choices from a variety of alternatives found within a large common stock of linguistic forms.

> The writer's 'choice' of available alternate forms is often determined by external conditions and then becomes the conscious, semiconscious, subconscious or (usually) unconscious result of habitually using one form instead of another. (McMenamin 2002, 164)

However, there are times when a writer has to consciously consider which forms to use since communicatively competent users are able to change their style of writing depending on the situation as they are aware that language is context sensitive (Hubbard 1995).

2.4 The idiolect debate

The notion of the idiolect is a central aspect of authorship attribution and one which has become quite contentious. In this section, I begin by analysing the nature of the idiolect

and how it is applied to authorship attribution, which is then practically exemplified with the example of the Unabomber. Thereafter I discuss the various objections to the notion of idiolect.

2.4.1 The idiolect and authorship attribution

'Idiolect' is defined by the *Oxford Concise Dictionary of Linguistics* as: "the speech or 'dialect' of an individual" (Matthews 1997, 169), which differs from 'dialect' which is defined as "any distinct variety of a language, especially one spoken in a specific part of a country or other geographical area" (Matthews 1997, 96). The idiolect is a central theme permeating authorship attribution.

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

Coulthard (2004) argues that every speaker of a language has over the years built up a substantial active vocabulary, which would differ from the vocabularies acquired by other speakers, and since everyone uses language differently "there will always be at least small differences in the grammar each person has internalised to speak, write and respond to other speakers and writers" (McMenamin 2002, 53). Idiolect can be summed up as the individual's unconscious and unique combination of linguistic knowledge, cognitive associations and extra-linguistic influences (McMenamin 2002).

Moreover, it is not only the size of the vocabulary that matters but the individual's preference for selecting certain items rather than others. Even though a speaker/writer could select any word at any time, they usually make their selection from a set of preferred lexical items. Even writers writing on the same topic can be expected to select a different set of lexico-grammatical items, even if they intend to express the same ideas. Coulthard (2004) uses an example of a disputed statement from the Appeal of

Robert Brown of how unique an utterance can be. On the surface, an utterance such as *I asked her if I could carry her bags* may not seem remarkable, but a Google search by Coulthard shows that it was in fact quite unique. I conducted the same Google test on the 22 September 2011 on this utterance and I found 9 instances. However, 5 of the 9 refer to Coulthard's research.

Figure 2.1 – Google search of "I asked her if I could carry her bags" (Coulthard 2007, 197)

String	**Instance**
I asked	2 170 000
I asked her	284 000
I asked her if	86 000
I asked her if I	10 400
I asked her if I could	7 770
I asked her if I could carry	7
I asked her if I could carry her	4
I asked her if I could carry her bags	0

This theory has led to the adoption of the somewhat unhelpful metaphor of the *linguistic fingerprint.* Coulthard (2004) explains that for forensic investigations of authorship attribution purposes, the idea of a forensic fingerprint is misleading, as it gives the illusion of enormous databases made up of huge numbers of linguistic samples of millions of idiolects, which could then be used to match and test a disputed text. Thankfully, it is highly improbable that a forensic linguist would be asked to identify a single author from millions of candidates on the basis of linguistic evidence. Olsson (2008, 31) asserts that: "the emphasis should probably be upon the relative difference between the candidate authors and how we can classify their texts". It would be prudent, when examining language style for authorship attribution purposes to rather consider what is distinctive, as opposed to unique (Olsson 2008).

2.4.2 Idiolect and the Unabomber

An example of Coulthard's (2004) argument of uniqueness of utterance can be found in the specifics of the Unabomber case, as a "persuasive example of the forensic significance of idiolectal co-selection" (Coulthard 2004, 432). In 1995, a 35,000 word manuscript entitled *Industrial Society and its Future* was written by an individual claiming to be the Unabomber who offered to cease his bombing campaign if his manuscript were to be published. In that document, he used the term: *cool-headed logician,* which was identified as Kaczynski's own terminology by his brother, who had read the manifesto. The FBI subsequently tracked down and arrested Kaczynski in his Montana cabin, wherein they found a number of documents which were later subjected to linguistic analysis. One significant document was a 300 word newspaper article written a decade earlier on the same topic. "The FBI analyst claimed major linguistic similarities between the 35,000 and 300 word documents: they shared a series of lexical and grammatical words and fixed phrases, which the FBI argued provided linguistic evidence of common authorship" (Coulthard 2004, 433). However, the FBI's analysis did not go unchallenged. The defence enlisted the services of a linguist who counter-claimed that it was not possible to attach significance to these shared items as anyone could use any word at any time, and for that reason shared vocabulary could have no diagnostic significance. The defence's linguist extracted twelve words and phrases (*at any rate, clearly, gotten, in practice, moreover, more or less, on the other hand, presumably, propaganda, thereabouts* as well as words derived from *argu* and *propos* such as *argument* and *proposition*), as examples of lexical items likely to appear in any polemical text. A subsequent Internet search conducted by the FBI revealed 3 million documents which included one or more of the twelve lexical items. Yet, once the search was narrowed to include all of the twelve lexical items only 69 were found that included examples of all 12 words and phrases, and all of those were versions of the 35,000 word manifesto (Coulthard 2004).

2.4.3 Objections to the importance of idiolect in authorship attribution

The Unabomber case highlighted just how useful the idiolect is in attributing authorship. However, the notion of idiolect as an important aspect of authorship determination has

not gone unchallenged, and in recent years a number of prominent forensic linguists, most notably Grant and Olsson, have questioned its validity. Grant (2010, 509) posits that “current debate within forensic authorship analysis has tended to polarise those who argue that analytical methods should reflect a strong cognitive theory of idiolect and others who see less of a need to look behind the stylistic variation of texts they are examining”.

Grant (2010) adduces that even if we accept the notion that “every native speaker has their own idiolect” (Coulthard 2004, 432) it is not necessarily true that an individual’s idiolect will be measurable in every text written by that individual, irrespective of the length of the text. Moreover, it would require a fairly substantial length of text before any measurable idiolect could be discerned, and in order to be useful in a forensic analysis, the idiolectal features would have to be repeated either in one text, or in a range of texts by the same author. Grant (2010) talks of the need to make a distinction between observation and theory when discussing idiolect. Even though the theory of an idiolect as distinctive variety of language may be necessary for authorship identification, the practical applications require the ability to detect consistent patterns of usage.

Cognitivist theories of idiolect adduce that language production is the result of linguistic competence, where linguistic performance is a reflection of an individual’s capacity to produce language (Grant 2010). The cognitive approach to authorship analysis takes the view that one can measure cognitive capacity, and therefore would employ tools such as measuring syntactic complexity. This method may yet prove to be of tremendous use to quantitative and computational linguists, who with longer texts are be able to mathematically describe observable authorship markers such as word frequencies and syntactic structures. However, it does not entirely explain consistency within an author, or distinctiveness between authors (Grant 2010)

Although sometimes viewed as being in opposition to the cognitivist theory of idiolect, stylistic theories of idiolect contend that understanding differences which occur between individuals is paramount (McMenamin 2002). Grant (2010) describes the stylistic theory

of idiolect as the interaction between habit and context, with the emphasis being on how and why language varies and/or stays constant within a sociolinguistic context. Grant (2010) asserts that this approach could offer a better explanation as to why there is variation between individuals than cognitivist approaches, as all individuals have different linguistic experiences, which will surface during their language production. Grant (2010) concedes that this is by no means idiolect-free analysis: the theory of the idiolect is unlikely to be abandoned. Instead, it could be viewed as a form of authorship analysis which draws on a different notion of idiolect, where the sociolinguistic context is taken into account along with the cognitive resources of the individual, which play an important part in formulating the person's idiolect.

2.5 Language variation

A feature of both dialect and idiolect is linguistic variation, where language used in both groups and individuals changes over time, or according to differing sociolinguistic contexts. The context of my study is concerned with the language used in digitally mediated communication, and more specifically Facebook, which has seen the evolution of new writing styles. This section will begin, firstly, by examining what language variation is, and why there is language variation and, secondly, how this variation affects authorship attribution.

2.5.1 Reasons for language variation

On the one hand, language variation and change which happens within a dialect or language can be viewed as a group phenomenon; for example, Crystal (2007) cites the example of how American English is becoming more prominent in speech communities which traditionally used the British variant of English. On the other hand, changes within an idiolect can be seen as a reflection of the individual's own use of language (McMenamin, 2002). As languages change and evolve over a period of time, there are periods when old and new forms will be used together in the whole speech community as well as in an individual. Johnstone (2000) states that through speech and other aspects of behaviour, individuals display their individuality as well as solidarity with their own social group.

Labov (1994) states that all languages demonstrate internal variation, caused by a variety of external factors, which can allow groups of speakers or writers, or individual speakers or writers to be differentiated from each other. This separation can be caused by time (different generations), geography, social factors (sex, age, gender, ethnicity, religion, income), and the immediate social context of the language use (topic, intended listener or reader). McMenamin (2002) describes the linguistic difference between two dialects in terms of their variable differences in pronunciation or spelling, word formation, sentence structure, word and sentence meaning and different ways of using the language. The linguistic changes which occur in languages, dialects or idiolects only really become apparent when compared to the accepted standard form of the language.

McMenamin (2002) describes linguistic variation as the presence of more than one way to say or write the same thing in the language of a community or an individual. Given that there are multiple forms of linguistic variants available to a language user, an individual or community of language users may always use a particular linguistic variant or may never use a particular linguistic variant. What is most common though is the relative use of two or more linguistic variables, where the linguistic variable is used to quantify the relative presence or absence of each variant against all possible occurrences of the variable. An example of this in digitally mediated communications is the writer's choice of using *you* or *u.* A writer may use both forms depending on the context, or may prefer to use only one form.

2.5.2 Language variation in authorship attribution

Olsson (2008) applies variation directly to a forensic context when he talks about *intra-author variation* and *inter-author variation.* Intra-author variation refers to the ways in which an author's text differs from another text written by the same author, whereas inter-author variation refers to the ways texts vary between different authors. Olsson (2008, 33-34) discusses eight different causes of intra-author variation, which have relevance when selecting texts for analysis, namely: (1) genre; (2) text type; (3) fiction vs non-fiction; (4) private vs public texts; (5) time lapse; (6) disguise; (7) changes in

circumstance; and (8) sociometric parameters. However, in my study, the only causes of variation that could have any bearing are *time lapse,* if some time has passed between posts, and *change in circumstances* if the writer has undergone any recent changes in her life. Despite Facebook being a social networking site, there could be *sociometric parameters,* as it is common for an individual to have on their friends list people who occupy different power positions, and that will affect the choice of language.

2.6 Language style

Halliday (1989, cited in Coulthard 2005, 9) states that written and spoken language are organised differently and that this can be seen in both the grammar and lexis. This section examines the notion of language style further, by firstly, looking at the differences between spoken and written language, and secondly, examining how this issue applies to DMC. Lastly, it discusses how linguistic norms are constructed, and then taken cognisance of that in authorship attribution, both from a stylistic perspective, where language features are described and compared, and from a stylometric perspective, where the features are counted and subjected to statistical analysis.

2.6.1 Spoken and written language

As a broad generalisation, spoken language uses shorter clauses and a lower ratio of lexical to grammatical words, whereas, written language uses longer clauses and has a higher lexical density (Coulthard 2005). Language style, therefore, needs to be discussed from both the spoken and written perspective.

> Style in spoken language relates to linguistic variation resulting from the social context of the interaction. The social context is defined by the topic and purpose of the interaction, as well as the social, cultural, and geographic characteristics of its speakers and listeners: their age, sex, race ethnicity, education, links to social networks, group affiliations, places of residence etc. (McMenamin 2002, 110)

Style in writing refers to the manner in which language is used in certain genres, periods and contexts. Just like spoken style, writing style also shares the social context

connection. Writers employ recurrent choices, usually the subconscious habit of choosing one form over another. Although McMenamin was not referring directly to DMC, it is possible to attach his notion of recurrent choices to DMC, for example: *you / u / ya* (Crystal 2007), where all three forms are considered correct, and depending on addressee, appropriate for DMC. McMenamin (2002, 110) talks of two types of choices: "variation within the norm and deviation from the norm". Variation within the norm means choices which conform to the norms of prescriptive grammar or which are considered correct. Deviation from the norm refers to choices which would be considered grammatically and/or lexically incorrect: *"*The norm itself must be defined in order for it to be used as the standard for identifying variation within it or deviation from it" (McMenamin 2002, 110).

2.6.2 Language use in DMC

Although features of DMC, which is referred to as *Netspeak* by Crystal (2007), often attempt to mimic speech, it is different from speech in its most fundamental properties. Crystal (2011, 17) states that speech is "time bound, dynamic and transient" and all the participants are present and the speaker has a definite audience in mind. Writing, on the other hand, is "space bound, static, and permanent" and the author is usually distant from the reader and may not even know who the reader/readers will be. In speech, unless deliberately initiated by the recipient, there is not any time lag between production and reception, which is in contrast to writing, where there is a definite time lag and writers must take this time lag into account, as well as knowing that their writing may be read and (mis)interpreted by numerous people. Millard (1996, 147) posits that in textual cyberspace "the linguistic and paralinguistic signs that maintain cognisance of the social relation between the sender and receiver of a message, are drastically reduced". This has a particularly noticeable effect when it comes to feedback and turn-taking and it is here where the interaction differs from conversational speech. During a face-to-face conversation the participants can utilise extralinguistic cues (facial expressions and gestures) to facilitate meaning, as well as employing deictic expressions such as *that one* and *in here* which refer directly to the situation at hand. However, writing does not allow for such nuanced meaning and there is no immediate

feedback for clarification. The language used on the Internet for social purposes has had to invent ways of expressing social functions in a written medium. Crystal (2011) offers three examples: (1) lexis is often characteristic of informal speech, particularly contracted forms (*isn't*); (2) coordinate sentences are frequently lengthy and quite complex; and (3) there is made-up vocabulary (*thingamajig*) and obscenity and slang are normal and may appear as a graphic euphemism (*f****). Despite this, it is quite difficult to represent prosodic features such as word stress, and as Crystal (2011, 19) states: "The many nuances of intonation, as well as contrasts of loudness, tempo, rhythm, pause, and other tones of voice, cannot be written down with much efficiency".

Crystal (2007) describes some of the inventive attempts at mimicking speech acts in DMC, for example, the use of upper case letters to show emphasis (*This is a VERY important point*) and pauses can be shown with dots (...). Due to the fact that Internet communications lack kinesics and proxemics, which are common in face-to-face communication, and are essential in moderating social relationships (Crystal 2011), emoticons were devised to reduce attitudinal ambiguity, where an individual emoticon, such as a basic smiley (☺), can express sympathy, delight or amusement. Research does seem to show that emoticons tend to be used predominantly by younger people, although older people seem quite prepared to use an emoticon to replace an entire utterance. Moreover, the use of emoticons appears to be more popular amongst females. In a study conducted by Katzman and Whitmer (1997) cited by Crystal (2011, 24), only one in six of the males used emoticons, as opposed to three quarters of the 16 female participants.

2.6.3 Linguistic norms

Norms can be viewed as either linguistic or statistical. Linguistic norms are further subdivided into, firstly, prescriptive, which refers to what is considered correct according to dictionaries and grammars and, secondly, descriptive, which refers to what the user considers appropriate use (McMenamin 2002). However, McMenamin (2002, 110) adds that "linguistic norms are not static; they evolve over time in a social, cultural, and geographic community of speakers and writers". Table 2.1 gives examples of

prescriptive and descriptive norms together with examples of what would be considered *the norm,* i.e. correct, acceptable variations within that norm, and deviations from the norm.

Table 2.1 – Examples of linguistic norms (McMenamin 2002,117)

Type of norm	**The norm**	**Variation within the norm**	**Deviation from the norm**
Prescriptive	*Examples*	*Examples*	*Examples*
Grammatically correct	I *am* going now.	I'*m* going now.	I *be* going now.
Socially appropriate	I'm afraid you're too late.	Sorry, the shop is closed.	Get the hell out of here!
Descriptive			
Prestige: US standard	We have enough money	We'*ve got* enough money	We *gots* enough money
Choice of variety: teenage	Hey, man!	Hey, dude!	Hello, Sir.
Class: age	That's a *cool* idea.	That's a *neat* idea.	That's a *swell* idea.
Regional: US dialects	A quarter to eleven.	15 *before - / of - / till* – eleven.	Eleven *less* fifteen.
Situational: at work	Where's the restroom?	Where's the bathroom?	Where can I take a leak?
Quantitative			
How often norms are used.	We *are* here. (10%)	We'*re* here (80%)	We here (10%)
In a defined social context.	It is me / It's me (85%)	It is I (10%)	It be me (5%)

The concept of norms varying over time and context is particularly relevant when considering how the language used in DMC has changed what is considered correct and appropriate. For example, prior to the advent of mobile telephony it would have been considered incorrect to use the lowercase '*i*' instead of the uppercase *'I'* when referring to the first person singular, yet now it is considered acceptable in DMC writing (Crystal 2007).

From a forensic linguistic perspective, it is worth noting McMenamin's (2002, 111) assertion that: "Prescriptive norms can be useful in authorship studies because they can be used to describe variation". A statistical norm describes the linguistic norm as a frequency distribution for each linguistic feature found within a linguistic community of either speakers or writers.

> Which forms speakers and writers of a community use can be counted vis-à-vis possible alternate forms, i.e., how often certain forms are used and in which specific linguistic and social circumstances. (McMenamin 2002, 116)

2.7 Style markers

McMenamin (2002, 172) raises the question: "How are style markers identified?" as being the most important issue in the current research on questioned authorship. This can be broken down into two distinct questions: "How are criteria for identification motivated, and how are stylistic variables selected and justified?". This section will analyse, firstly, the question of what constitutes a suitable style marker and how they should be chosen. Secondly, there will be a review of some of the reservations expressed regarding style markers, along with some cautions to be kept in mind. Lastly, there is a discussion of research done using some of the most important style markers that have been used in my study, namely: idiosyncrasies of punctuation; typography; spelling; lexis and DMC features on the stylistic side, and keyness; function words; most frequently occurring words, and overall punctuation mark frequencies on the stylometric side.

2.7.1 What constitutes a suitable style marker?

Despite the long history of authorship attribution, there is still doubt about what constitutes a reliable authorship marker and how to identify one, especially within a forensic linguistic context where short texts and small samples are the norm (Grant and Baker, 2001). According to Rudman (1998), there are at least a thousand style markers which exist in stylometric research. However, he has since updated that number to a figure in the millions, particularly with the aid of the computer program *Docuscope* (J.

Rudman, pers. comm). McMenamin (2002, 216 – 231) offers a very useful list of style markers, which has been employed in over eighty cases. The style markers can be categorised as character-based, word-based, sentence-based, document based, structural or syntactic. A few examples of style markers include: function word usage (common adverbs, auxiliary verbs, conjunctions, prepositions and pronouns); word collocations; sentence length and punctuation. However, Baayen et al. (2000) point out that style markers may still be sensitive to differences in genre and topic, especially when the text corpus is small. This is particularly true when applying stylometry to e-mail authorship, instant messaging and social networking communications, particularly since these forms of communication tend to be very short. Grant and Baker (2001) discuss the characteristics of a good style marker and how it can be identified without falling into the trap of generalising. Since authorship attribution is a classification problem, it leads to the conundrum of: “What stylistic features can discriminate between these texts by different authors?” (Grant and Baker 2001, 68). Therefore, considering the almost impossible task of finding valid and reliable style markers that would be applicable to all writers, due to the inherent variability of language, irrespective of whether that variation is dialectal or idiolectal, it would be prudent to utilise an array of style markers which would consist of those markers which collectively account for the most variance in the text (Grant and Baker, 2001).

2.7.2 Cautions regarding style markers

Olsson (2008) describes two opposing views relating to style markers. On the one hand, style markers are consciously chosen by an individual and can be observed and measured. On the other hand, style markers are unconscious habits not controlled by an individual, but once discovered by a linguist, they can be observed and measured.

> The main assumption underlying stylometric studies is that authors have an unconscious as well as a conscious aspect to their style. Every author’s style is thought to have certain features that are independent of the author’s will, and since these styles cannot be consciously manipulated by the author, they are

> considered to provide the most reliable data for a stylometric study. (Holmes 1997, 2)

This dichotomy raises a number of issues: If style markers are conscious habits then it stands to reason that an individual can alter his or her use of style markers and they can even be imitated by a third party. If, on the other hand, style markers are, in fact, unconscious habits then it needs to be determined whether the style markers differ from individual to individual or used identically by all writers and speakers (Olsson, 2008). From a practical forensic position regarding the use of style markers, Olsson (2008) issues the following warning:

> There are several important points to be noted about style markers. First, to measure unconscious style markers meaningfully, you need a great deal of text – such as a full length novel, or hundreds of short texts. On the other hand, the fact that we can observe certain style markers tells us that they are open to imitation – unless we are able to demonstrate that there is some kind of systematic or structural link between them. (Olsson 2008, 29)

Olsson's (2008) assertion that one needs extraordinarily long texts appears to be at odds with other researchers in the field of forensic linguistics. Chaski (2011), in discussing the case of *Ceglia v Zuckerberg* cited by McMenamin (2011) on the *Language Log* Internet forum, states: "I have also tested for minimal data requirements, and have found that 2000 words and/or 100 sentences per author affords the most robust results". However, all researchers would agree that more text is definitely preferable to less text, and in my study I will investigate just how much difference a doubling from 1000 to 2000 words makes in the accuracy of the authorship attribution.

2.7.3 Some style markers related to the study

The following style markers: keyness, function words, punctuation and spelling, lexis and most frequently occurring words will be described in more depth as these style markers will be employed in the study.

2.7.3.1 Keywords

The analysis of keywords is concerned with lexical items whose usage is unusually frequent or infrequent when compared to a reference corpus (Bednarek 2009).

> A word is said to be *key* if [...] its frequency in the text when compared with its frequency in a reference corpus is such that the statistical probability as computed by an appropriate procedure is smaller than or equal to a *p* value specified by the user. (Scott 2012, 174)

The *p* value, which is usually 0.05, is specified by the researcher, and using either the log-likelihood or Chi-square statistical test, it can be determined whether that *p* value is met, thus determining whether the difference is significant or not. The results of the log-likelihood or Chi-square statistical test give the keyword its degree of keyness (Gabrielatos and Marchi 2011): in other words, keyness is a measure of significant difference.

Kotzé (2010) was the first researcher to employ the keyword function of *WordSmith Tools* in a forensic linguistic case in South African legal history. The case he worked on was the State vs. Kerr Hoho, commonly referred to as the Father Punch case, a High Court matter in the Eastern Cape. In this case, the accused was convicted on a number of charges of criminal defamation due to the publication of a series of documents written under the *nom-de plume* of Father Punch. These publications contained accusations of bribery, financial embezzlement and corruption, allegedly perpetrated by high ranking politicians in the Eastern Cape and national government. Kotzé (2010) started by dividing the documents to be assessed into two groups. The first was a set of memoranda, (5,482 words) known to have been written by the accused to the Eastern Cape Provincial Legislature and other groups (the “core” document, which would also serve as the reference corpus). The second set of documents was eleven “chronicles”, (25,431 words) that were thought to have been written by the accused (Kotzé 2010, 188)

Kotzé (2010) found, by using the keyword function of *WordSmith Tools,* that If the two texts were written by different authors, one could usually expect a higher keyness value and fewer content words and that the higher the keyness, the more unlikely the two texts were to have been penned by the same author. Table 2.2 shows an example of a keyness analysis of two different authors.

Table 2.2: Comparison of keyness between texts by different authors (Kotzé 2010, 189)

Word	**Keyness**	***P* - value**
Of	45.5	0.000000
In	35.4	0.000000
He	33.1	0.000000
His	32.2	0.000000
Which	31.9	0.000000
Was	27.7	0.000000
More	25.2	0.000001
Can	22.9	0.000002
About	19.9	0.000008
Is	13.1	0.000298
Average	**28.7**	

Kotzé (2010) determined that in the analysis of keyness between texts written by the same author the keyness value was quite low and that the lower the keyness values, particularly with grammatical words, the higher the likelihood of common authorship. It was suggested that a keyness analysis which exhibits a high ratio of content keywords, with a low keyness value would reinforce the probability of common authorship, as content words speak directly to the author's idiolect and the context (Kotzé, 2010). Table 2.3 shows an example of a keyness analysis of texts that were subsequently found by the court to have been written by the same author.

Figure 2.3: Comparison of keyness between the core document and anonymous chronicle
(Kotzé 2010, 191)

Word	**Keyness**	***P* - value**
Of	6.4	0.011519
Back	5.9	0.014838
Is	5.8	0.015915
Should	4.0	0.044733
Does	3.1	0.078846
Gqobana	3.1	0.078846
Person	3.1	0.078846
Same	3.1	0.078846
Smith	3.1	0.078846
Money	3.0	0.085371
Average	**4.1**	

Kotzé (2010) notes that computer programs such as *WordSmith Tools* are unable to differentiate between function (or grammatical) words and content words, and in many cases the difference is not always obvious and can be dependent on the context. Keywords are considered good indicators of what the text is about, as the keywords will be operating within a restricted range of topics (Scott, 2012), and this would be particularly true of content keywords. Thus it is possible that there could be significant discrepancies between the writings of the same author if content keywords are considered, particularly if the writings are in different genres. The case for function keywords is somewhat different, where differences in function word usage would suggest dissimilar “grammatical vocabularies” as function words vary far less in density across the writings of a single author than lexical words (Kotzé 2010, 188). This work highlights the importance of keywords, and especially grammatical keywords, in identifying the author of disputed texts. These keyword analyses, in conjunction with stylistic analyses, were sufficient to satisfy the court.

2.7.3.2 Function words

Like Kotzé, Olsson (2009b) describes two types of words: lexical words and function words. Lexical words include word classes such as nouns, verbs, adjectives and certain adverbs. Function words are “semantically bleached” (Juola 2006, 265), meaning they have little or no independent meaning, but instead, they carry the grammar of the

language and include prepositions, determiners and function adverbs. Morton (1978, 133) offers a list of 23 of the most commonly occurring function words, namely: *a, all, also, and, any, as, at, been, but, for, in, it, no not, of, on, so, that, that, the, this, to, very* and *was.* However, Morton's list was drawn up to analyse literary texts rather than social writing between friends on Facebook, and although "high frequency function words are relatively independent of genre", one still has to be cognisant of the effects of genre (Hubbard 1995, 61). For example: in a study conducted by Hubbard (1995), which involved threatening letters being compared to expository writing, it was argued that the nature of extortion letters lent itself to the use of *no* in warnings and threats (*there will be **no** escape, If offer rejected or ignored and **no** message left*), which is less usual in expository and narrative texts.

Olsson (2009a, 98) compared the frequencies of the definite article (*the*) and indefinite articles (*a/an*) across two separate genres of writing: news articles and e-mails.
Table 2.4 shows the results.

Table 2.4: Distribution of <u>the</u> and <u>a/an</u> across a corpus of news articles and e-mail
(Olsson 2009a, 98)

Word	News articles	E-mail
the	0.074	0.044
a/an	0.027	0.026

It can seen that the distribution for *a/an* is similar across the two genres whereas the difference for *the* is quite significant. This skewing effect due to genre differences has a sound linguistic basis, as news articles are usually about something as opposed to e-mail (and Facebook messages), which are more about *you* and *I*. Hubbard's (1995) and Olsson's (2009a) studies highlight the need for caution when using function words across different genre types.

> If comparing texts of different types then we should not rely on frequency counts of some common function words for authorship purposes: the text type and genre influences are likely to skew the result. (Olsson 2009a, 98).

Research conducted by Olsson (2008) highlights the usefulness as well as the pitfalls inherent in using function words as a style marker. Olsson (2008) states that *the* is the most commonly used word in English. Among L1 English speakers it is used on average once every 20 words in both speaking and writing. Olsson (2008) believes that it has very little value as a determiner of authorship on its own across all text types, yet if genre and text type are controlled, it may have value as an authorship marker. However, it would be highly unusual to rely on a single function word as a style marker. Rather, one would utilise a range of function words, such as Morton's (1978) list or, alternatively, function words mined from the text, in order to build a profile of the writer, which could then be analysed against the profile of another writer, as exemplified in Hubbard (1995).

2.7.3.3 Punctuation and spelling

Chaski (2001) describes two style markers: punctuation and spelling, which would be particularly pertinent to a study of DMC, since writing on social networking sites is very encouraging of creative spelling and punctuation. A few ways to analyse the use and non-use of punctuation marks are to count the frequency of use within a text, look at where the punctuation marks are used and whether the author has any idiosyncratic uses of punctuation marks. However, Chaski (2005, 5) adds to this by stating that punctuation "has only really been successful when combined on its own with an understanding of its syntactic role in a text". Olsson (2008) concurs by saying that analysing punctuation is successful because of what the punctuation marks are doing in a sentence. If one uses the comma as an example: it divides clauses, separates noun phrases and signals a break before or after a conjunction. Punctuation is particularly useful when dealing with short texts as it is highly probable that the number of punctuation devices will be more than any single word, and they are likely to occur in sufficient quantities to be statistically counted (Olsson 2006).

> In simple punctuation approaches, the punctuation marks themselves, such as commas, colons, exclamation points, etc., are counted as being sentential, clausal, phrasal, appositive or word internal. In the syntactically classified punctuation approach, the marks (no matter what they specifically are) are counted by the kind of boundary or edge which the punctuation is marking. (Chaski 2005, 5)

However, Grieve (2005) points out that there have been surprisingly few attribution studies based primarily on punctuation, probably due to sentence length having been rejected as an indicator of authorship. However, in modern texts, particularly those in DMC, with its creative uses of punctuation (Crystal 2011), there is "a great deal of optionality in how an author chooses to use these grammatical characters" (Grieve 2005, 19). Crystal (2011) describes how, in DMC, dashes are used to show a change in direction of thought, the use of dots to express incompleteness and commas are used to show pauses in rhythm. Moreover, when dealing with DMC, it is necessary to move beyond the traditional set of punctuation marks (full stops, commas, question marks etc), to include symbols such as #, @ and carets (^) as well as emoticons, such as smilies (☺), which may perform punctuation duty (Crystal 2011).

In the legal dispute between Facebook founder Mark Zuckerberg and Paul Ceglia over Ceglia's claim to part ownership of Facebook in July 2011, Professor Gerald McMenamin was asked to analyse known Zuckerburg e-mails against questioned e-mails purportedly from Zuckerberg to Ceglia. McMenamin's report showed that he had analysed 11 style markers, and of those 11, three were spelling and two were punctuation, namely apostrophes and suspension points (ellipsis). In the questioned Zuckerberg texts, there appears to be a number of errors regarding apostrophes: *doesnt, parents* (meaning *parents'*), *sites* (*site's* = contraction for *site is*) and *sites* (*site's* = possession), whereas in the known Zuckerberg texts all contractions and possessives are used correctly. The second punctuation style marker that was analysed was suspension points. In the questioned text, there is one example of suspension points and the points are spaced (***. . .*** *I've been tweaking the search engine today*), whereas

in the known Zuckerberg texts there are three examples of suspension points and they are not spaced (1) (*online as quickly as I can* **...**), (2) (*So let me know* **...**), (3) (*boxes* **...** *there*) (McMenamin 2011).

Although spelling errors, alternative spellings and idiosyncratic spelling are a common feature of DMC (Crystal 2007 and 2011) the literature urges caution when applying spelling to authorship attribution. Chaski (2001) states that it is commonly believed that spelling errors are unique to individuals and constant in an individual's writing, when, in fact, people's spelling habits can change over time due to education or being exposed to different forms. Furthermore, Goutsos (1995) adds that spelling errors are not so odd that they cannot be shared and that different writers can and do produce the same errors. One high profile case which highlights the difficulties in using spelling as an authorship attribution marker is that of the Unabomber (Grieve 2005). The sociolinguist, Robin Lakoff, acting for the defence, objected to the FBI's James Fitzgerald's reliance on spelling evidence. Fitzgerald used the argument that both Kaczynski and the Unabomber made use of British spellings (e.g. *analyse* and *licence*), but Lakoff argued that they were not unusual enough in American English to show that Kaczynski was in fact the Unabomber. Lakoff also submitted spelling based counter evidence wherein she contended that the Unabomber had spelled the word *chlorate* correctly, unlike Kaczynski, who rendered it as *clorate.* Foster's counter highlights the basic problem with error-based authorship attribution. Kaczynski may not have spelled *chlorate* properly, but in later texts he had spelled *chloride* and *chlorine* correctly. Foster's argument was that if he had learned to spell *chloride* and *chlorine* correctly at the time the Unabomber texts were published, then he should have grasped the correct spelling of *chlorate*.

Referring back to the aforementioned Zuckerberg/Ceglia case, McMenamin (2011) also analysed three discrete spelling examples. *Backend* and *frontend* are technical terms which were rendered as two words in the questioned text (*the back end of the site*) and as one word multiple times in the known Zuckerberg texts: *backend* (6x) and *frontend* (5x). In the questioned text the word *Internet* begins with a lowercase *i* and in the known

Zuckerberg texts both examples of *Internet* begin with a capital *I.* Lastly, the questioned text has two examples of *can not* as two words, whereas the known Zuckerberg texts has six examples of *cannot* written as one word (McMenamin, 2011). However, it would be fair to note that this case, and its conclusions have been criticised by various prominent academics in the forensic linguistics field. Dr Ron Butters asked whether there had been sufficient evidence, whether the evidence had been suitable, and whether it had been evaluated in an appropriate and convincing manner. Butters as well as Chaski argue that the numbers were too small to be statistically testable (Liberman 2011).

A further consideration regarding spelling, particularly when dealing with forensic cases, is the ease of deliberate obfuscation. Leonard (2005) presents the following example of a kidnap case investigated by Dr Roger Shuy in 2001. The pencil scrawled note read as follows:

> *Do you want to see your precious little girl again? Put $10 000 cash in a diaper bag. Put it in the green trash kan on the devil strip at corner 18th and Carlson. Don't bring anybody along. No kops!! Come alone! I'll be watching you all the time. Anyone with you, deal is off and dautter is dead!!*

Shuy noticed the obvious spelling mistakes: *kan, kops* and *dautter* which are juxtaposed to the correct spellings of far more complex words: *precious* and *diaper,* as well as fairly standard sentence structure and punctuation. This led Shuy to the conclusion that the kidnapper was fairly educated and attempting to appear uneducated (Leonard 2005). Moreover, this case highlighted the fact that deliberate obfuscation can, sometimes, be easily identified.

From the perspective of DMC, the extent of the relevance of spelling would depend on the medium. For example, most e-mail programs have a spell checker that would have highlighted *kan, kops* and *dautter,* whereas social networking sites such as Facebook

and Twitter, as well as mobile phones, with the exception of smartphones, do not have any spell checkers.

2.7.3.4 Most frequently occurring words

Juola (2006, 262) states that: “The simplest way to confirm or refute authorship is simply to look for something that completely settles the authorship question” such as a word which only occurs once and is quite distinctive. Shuy’s example of *devil strip* in the ransom note (above) is an excellent example of *hapax legomena,* or a word which occurs only once in a text. In the note, the kidnapper refers to a *devil strip,* which is the grass strip between the pavement and the road. Unfortunately for the kidnapper, this word appears to be only used in Akron, Ohio, and it is relatively unheard of, even in nearby Cleveland. Since the police had only one suspect from Akron on their shortlist, he was arrested and charged (Leonard 2005). The above example highlights how an individual word can offer strong clues as to the author’s group identity. For example, if an author were to write about sitting on a *chesterfield,* then it would be assumed that the author was not only Canadian but an older Canadian (Juola 2008). However, there is a very serious concern regarding this sort of analysis, and that is that it is easy to fake. Olsson (2010) describes the case of Peter Chapman, known as the Facebook murderer, who was able to manipulate his texting language to appear as both a 17 year old boy and his father. My study took a more cautious view of the value of uniquely occurring lexical items as being examples of *hapax legomena,* but covered similar ground by focusing in the stylistic investigation on non-standard lexis such as *famdamilies.* The stylometric analysis focussed on the most frequently occurring words.

When lexical preference and commonly occurring words are used in authorship attribution, it is on the assumption that the frequencies of words in the text are a direct function of the author’s lexicon (Grieve 2005). However, it should be noted that there has been very little research conducted into lexical preference in authorship attribution.

> The basic assumption is that the writer has available a certain stock of words, some of which he/she may favour more than others. If we sample a text produced by that person, we might expect the extent of his/her vocabulary to be reflected in the sample frequency profile. If, furthermore, we find a single measure which is a function of all the vocabulary frequencies and which adequately characterises the sample frequency distribution we may then use that measure for comparative purposes. (Holmes 1985, 334)

However, the lexical choice of a text is influenced more by the subject matter than the author. Even though every lexical item will be a product of the author's lexicon, different subjects will require different vocabulary, and not all sections of an author's lexicon will be equally rich. For example, if an author is very knowledgeable about a subject, it stands to reason that he or she will employ a larger, and more varied lexicon than on a subject on where the author's knowledge of the subject matter is limited.

2.8 Conclusion

The aim of this literature review was to bring together research conducted in stylistics, stylometrics, forensic linguistics and DMC, which was relevant to my study of authorship attribution on Facebook. Firstly, I situated forensic linguistics, with its sub-fields of stylistics and stylometrics, into its historical contexts. Unlike traditional stylometric cases, which focussed largely on literary work and involved a great deal of text, modern investigators of criminal activity have to work with far shorter texts. Moreover, mistakes made in authorship attribution of literary work would not have nearly as important consequences as those made in court cases, and so forensic linguistic standards have to be very high.

It is generally accepted that each individual uses language differently as a result of numerous sociocultural influences. It is these differences in language use that are of particular interest to an authorship attribution study. Therefore, a central thread running through forensic linguistics is the principle of idiolect, the idea that each person has their own unique version of the language that they speak, and these differences can be

observed and in certain cases statistically analysed. Despite some disagreement amongst scholars regarding the exact nature of idiolect, it remains an essential aspect of authorship attribution. Yet an individual's language use is not static: people vary the way they speak and write, whether consciously or subconsciously, and for a variety of reasons, such as the context of the interaction or the relationship between the participants, and even as a result of growing older.

Writers on social networking sites have a great deal of choice over which forms to use, riding the continuum from the accepted standard norms on the one end to the extreme colloquial forms associated with DMC on the other. This relatively new form of language exhibited in DMC on social networking sites such as Facebook has evolved to become a form of written speech (Crystal 2007 and 2011), where people generally write as though they were having a conversation with the other participant or participants.

This leaves the question: what does one need in practice to attribute authorship on a social networking site? In order to get as accurate a result as possible, it is best not to rely on just one type of style marker. Therefore, this study examined the use of a number of different style markers, which had all been used by different researchers in different mediums to social networking sites. It started with analysing keywords as championed by Kotzé (2010), which are useful in analysing which words are used unusually frequently in infrequently across two authors. The second set of style markers to be examined was the 23 function words proposed by Morton (1978). After that, the review examined using punctuation and spelling, as discussed by Chaski (2006). Spelling and punctuation are particularly relevant to research into authorship attribution on a social networking site, as idiosyncratic use of such features is not only tolerated, but encouraged. Lastly, drawing on research by Grieve (2005) and Juola (2006), there was discussion of analysis of the most frequently occurring words.

In Chapter 3, the theoretical aspects discussed in this literature review will be linked to practical use in my study, as its methodology is examined.

Chapter 3 – Research method

> *The general idea is that we can formulate our research hypothesis and eventually draw any necessary inferences or explain the phenomena we have observed.*
>
> (Olsson 2004, 45)

3.1 Introduction

This research method chapter begins by examining some ethical considerations related to this study and then moves on to discuss how the participants were chosen. The next section describes how the data was collected, followed by a review of the research methods employed, with reference to similar research conducted by investigators in the field of forensic linguistics. The research methods and procedures are informed largely by the frameworks offered by Hubbard (1995), Chaski (2001 and 2005), McMenamin (2002), Olsson (2004 and 2008) and Kotzé (2010).

The chapter then focuses on the qualitative analysis and examines the importance of stylistic analysis in a forensic linguistic context, as opposed to relying solely on a computer processed quantitative analysis. This part of the chapter examines the method used to identify the features in the participants' writings that make each set of writing unique. My primary focus is on Writer X's writing, as that is my disputed text, and all other texts have to be referenced against it. The stylistic analysis involves analysing the texts for punctuation, spelling, typography and word choice, which would be considered unusual or idiosyncratic, and then comparing them individually to the disputed text. Even though features such as punctuation marks can also be subjected to a stylometric analysis, this section will focus on the stylistic approaches in the analysis of those features.

The third main part of this chapter deals with the quantitative analysis and examines the importance of stylometric analysis within the field of forensic linguistics. It starts by looking at the workings of the concordance program *WordSmith Tools* (*WST*), before moving on to a discussion on the interpretation of the statistics used and a description of how the Chi-square test is used in forensic linguistic studies. Thereafter, there is a

review of how the style markers chosen for the stylometric analysis (keywords, function words, most frequently occurring words and punctuation) are analysed using *WST* at both the 1,000-word and 2,000-word level.

3.2 Ethical considerations

All the participants in the study were made aware that participation was purely voluntary. The voluntary nature of participation was made explicit to the participants through the process of informed consent (see Appendix 1 for consent letters). Participants were made aware that they could withdraw at any time and that no one would be advantaged or disadvantaged by participating or not participating. Confidentiality has been assured by replacing all names in the submissions with the word *name* in italics and replacing telephone numbers with either *cellular number* or *telephone number* and e-mail and Skype addresses were replaced with the phrases *e-mail address* or *Skype address*. Moreover, all screenshots showing names and faces have been blurred.

3.3 The participants in the study

The eight participants for this study were matched as closely as possible according to sociolinguistically pertinent demographics such as age, sex, race, dialect (sociolect), class and education level. They fell into the following demographic category: they were all female, aged between 30 and 40, educated to post-matric level in South Africa, spoke English as a home language and were active users of social networking sites and had been so for quite some time. All the participants' names were removed and pseudomised, as A through to H, with the disputed text being referred to as X.

Being confined to such a tight demographic meant that the sample of participants should have fairly similar writing styles, which helps us to avoid any linguistic separation (Labov 1994) caused by differences in social demographics.

> Since dialectal features are relatively well-documented and easy to spot, the more difficult case is differentiating among documents from the same dialect. If the tested technique can differentiate authors of documents which share dialectal

> features, then it can certainly work on documents which do not share dialectal features. (Chaski 2001, 4)

Therefore, if it can be determined that it is possible to distinguish different authorship styles within a relatively homogeneous group, it should, therefore, be feasible to distinguish authors in a more heterogeneous group. However, due to the fact that there is such a narrow focus, it may not be possible to extrapolate the findings to other homogeneous groups.

I am fortunate to have been given access to writings from such a homogenous demographic group, as other researchers have had to work with participants from different demographic groups rather than from one homogenous group. In Hubbard's (1995) extortion letters case, the suspect was a Romanian born engineer with Polish as a home language. Some of the comparative texts were written by speakers of Polish, who were qualified engineers. In Chaski's (2001) study, where she tested a number of authorship attribution hypotheses, she made use of a participant group made up of different ages and races and both genders.

3.4 Data collection

Chaski (2001) states that the parameters of a genuine case have to determine the design of a simulated case, and the task in all empirical tests is to distinguish between different writers, and to identify texts by the same writer, some of which are known and one is unknown. Secondly, the known writing samples must be selected on the basis of demographic homogeneity to ensure they qualify as 'suspects'. By choosing known writers who share sociolectal or dialectal features, it is possible to test for idiolectal rather than sociolectal or dialectal linguistic performance.

In order to obtain the texts I needed, I mined my Facebook friends list for candidates who met the demographic criteria. My next step was to send a Facebook message (Appendix 2) to all the potential candidates. Initially, I had intended to analyse 16 texts, eight females and eight males. Unfortunately, I was unable to obtain sufficient data from the men. Therefore, this study focused only on the females' texts. I received 15

responses, which I narrowed down to eight. The chosen eight were the only ones that could provide the necessary 2,000 words and met the stringent demographic criteria. My request to them was for 2,000 words of text from their Facebook inbox, cut and pasted onto a word document and e-mailed to me. The text had to be their own writing with no third party submissions and no editing of the text before submitting it. I asked them to start at their latest text and move backwards until 2,000 words had been obtained so as to avoid any significant time lags (Olsson 2008), and to have writings which were as current as possible. This also helped ensure that all submissions were from the same period of time. I asked one of the participants to submit an extra 2,000 words which would act as the disputed text X. Ideally, it would have been better if I did not know who the writer of the disputed text was, but it would have been, logistically, very difficult to arrange, as they do not all know each other. I did not alter the content of the texts, apart from changing names and removing all contact numbers and electronic addresses. Appendix 3 contains the full submitted texts from the eight participants.

3.5 Research methods

McMenamin (2002) and Olsson (2008) both offer a framework for an authorship comparison study, and I use a combination of these to conduct my own research. Step 1 is to assemble all questioned and known writings and check for compatibility and comparability.

Step 2 is to stylistically analyse the texts for similarity and differences, keeping in mind that it is important to include any counter-examples (Olsson 2008). I examine the questioned writing, looking for features which could be considered idiolectal to the relevant writer and list them, and then examine the known texts for the features found in the questioned text, as well as apparently idiosyncratic features which are not found in the questioned text. Features will be chosen using McMenamin's (2002, 120) criteria: (1) deviations from any norm such as errors or mistakes; and (2) variation within the writer's norm (i.e. does the writer use more than one form in a text (*u*/*ya*/you)).

Step 3 is the quantitative analysis, where I use the concordance program *WordSmith Tools* (*WST)* to count the frequencies of function words and punctuation marks and to look for keywords, followed by a Chi-square test of the profiles comparing the questioned text to each of the known texts.

Wachal (1966, cited in McMenamin 2002, 117) describes three models of authorship analysis: resemblance, consistency and population. The resemblance model is used when the number of probable authors has been limited to just one or a small number of authors, “the authorship question is defined narrowly to exclude or identify just one suspect writer”. This method is often employed when dealing with disputed wills. The consistency model is “used to determine whether two or more writings were written by the same author” McMenamin (2002, 118). This method is often used when dealing with multiple questioned letters which have allegedly been signed and written by different people and the people involved deny having written the texts. It becomes necessary to determine whether those texts were written by one or more authors. This method is also used when a single person claims responsibility for a text, yet the content and/or style of writing suggest the possibility of numerous authors, as was the case with the Federalist papers (Olsson 2008). Finally, the population model is employed when there is not any nonlinguistic evidence pointing to any candidate authors. A text then has to be analysed against a population of potential authors. McMenamin (2002) presents the example of a large insurance company which had received incriminating letters pertaining to one of its regional managers. The incriminating letters had to be compared to the known writings of 17 disgruntled and former employees. However, these models of authorship analysis are seldom used in isolation and are frequently used in combination. In my research, I begin with a population model as I compare eight writers against a disputed text and the various tests I will administer will exclude those whose profiles are completely dissimilar to the disputed text. Thereafter, the resemblance model will be used to exclude the remaining writers until the writer with the closest profile is identified. McMenamin (2002, 119) concludes by stating that “no matter what model is used to formulate the research question in a case, the analytical tasks of observation, discovery,

and comparison and contrast of style markers in separate sets of questioned and known writing is the same".

McMenamin (2002), as noted above, describes two approaches to authorship identification: qualitative and quantitative. Qualitative analysis involves isolating features of a text, and then identifying and describing those features as being characteristic of a single writer. A quantitative analysis involves isolating a feature and then measuring that feature in some way, for example, measuring its relative frequency in a given set of texts.

> Qualitative and quantitative methods complement one another and are often used together to identify, describe, and measure the presence or absence of style markers in questioned and known writings. (McMenamin 2002, 76)

The submitted texts in my study are subjected to both qualitative and quantitative analyses, with the first being the qualitative analysis.

3.6 Qualitative analysis

McMenamin (2002) states that describing the language of a relevant text should be the first step to analysing and interpreting the text. The first part of this section will look at issues relevant to the stylistic analysis of this study, and the second part will show how the qualitative analysis is conducted.

3.6.1 Aspects of a qualitative analysis

A qualitative study looks at *what* forms are used and *how* and *why* they are used (Johnstone 2000, cited in McMenamin 2002,129). However, qualitative analysis within forensic linguistics has been criticised for not being sufficiently scientific. Chaski (2005, 2) argues that "Without the databases to ground the significance of stylistic features, the examiner's intuition about the significance of stylistic features can lead to methodological subjectivity and bias". McMenamin (2002, 129) concedes that a qualitative analysis on its own will not achieve an "absolute conclusion about any kind of

indirect evidence, like a set of known and questioned writing". Yet, despite that, a qualitative assessment is still relevant for the following three reasons: (1) a qualitative analysis is the first step in order to discover, describe and categorise relevant linguistic features within a text; (2) qualitative evidence is far more demonstrable in a court of law than quantitative evidence, particularly if it precedes that quantitative evidence; and (3) the nonmathematical nature of qualitative analysis "will appeal to the structured sense of probability held by judges and juries" (McMenamin 2002, 129). Kotzé (2010) noted that:

> The technical nature of a purely statistical analysis would not necessarily be sufficiently transparent for a presiding officer (or jury) to come to an informed conclusion. It was found, in this case, that the combination of the qualitative analysis (which highlighted the correspondences) and the graphical representations of the statistical findings (which showed that the differences between the texts were largely insignificant) addressed precisely this aspect.
> (Kotzé 2010, 191)

Qualitative data can be collected and evaluated using numerous different methods and it does not have to be numerical, and even if it is numerical, it is not always necessary or possible to conduct a statistical evaluation (Olsson, 2004). This is particularly true when analysing texts from digitally mediated communication where the frequency of features may be too small, as could be seen from McMenamin's work on the Zuckerberg/Ceglia case discussed in Chapter 2, which was criticised for being 'unscientific'.

> While linguistic data frequently present countable variables, sometimes the linguistic significance of an identified variable is not captured by counting, or a variable is linguistically significant, but it does not occur regularly enough to be meaningfully counted. (Cohen 1977, cited in McMenamin 2002, 131)

3.6.2 Markedness in qualitative analysis

Markedness has its origins in structuralist and generative grammar as a means to explain constraints on grammatical rules. Within the field of forensic linguistics, markedness is relevant when looking for unique features of a text, features which are unusual or 'non-standard' (Olsson 2004). In this case unusual and 'non-standard' are different concepts, despite their superficial similarities. Unusual refers to a usage which does not conform to a general pattern: for example, the American spelling *color* in a text by a person who generally uses British/South African English. Non-standard refers to usage which is not in line with prescriptive grammar: for instance, the use of lowercase *i* when referring to the first person singular pronoun. Olsson (2004, 56) offers the following order-of-importance which is relevant when qualitatively assessing a text with markedness in mind: (1) types of grammatical structures, (2) punctuation, (3) idiom, (4) spelling and (5) document layout.

3.6.3 Mistakes and errors

When evaluating features such as spelling and punctuation, it is important to determine whether the unusual or 'non-standard' feature is the result of an *error* or *mistake*, which is somewhat difficult in practice. Both *mistakes* and *errors* relate to deviations from the standard norm. Mistakes are examples such as the mistyping of *the* as *teh* and *ing* as *ign* where the writer could subsequently recognise that they have deviated from the norm and will, if noticed, make the corrections. Errors are, on the other hand, the result of an author having acquired a different rule from the standard norm; for example, this author until quite recently wrote *publicly* as *publickly.* When analysing mistakes and errors in a qualitative assessment of a text it is necessary to look for consistency – does the author repeat the same mistake habitually (Coulthard 2007). This was exemplified in the Zuckerberg/Ceglia case, as analysed by McMenamin (2011), where Zuckerberg consistently wrote *cannot,* and in the questioned text it was written as *can not*.

3.7 Qualitative assessment of the participants' writing

This section describes the practical procedures followed in the qualitative analysis. It starts with how the features were identified and categorised, and then moves on to discuss the descriptors used to describe the degree of difference between the authors.

3.7.1 Categorisation of stylistic features

The stylistic analysis of each participant's writings is divided into two parts. Firstly, noticeable features are extracted and put into tabular form (See Table 3.1). The table has two distinct parts. The first shows the stylistic features shared with the writer of the disputed text (X). Therefore, the shared features are not the same for all the writers. Even though this is a qualitative analysis, frequency counts are provided where relevant. Frequencies marked with an asterisk show features with statistically significant differences. The frequency column is divided into four separate columns, the first two columns (C1 and X1 in Table 3.1) give the frequency counts for the 1,000-word level, and last two columns (C2 and X2 in Table 3.1) give the frequency counts for the 2,000-word level.

Table 3.1 – Stylistic analysis of Writer C

Features shared with Suspect X's text	**Examples**	**Frequency**			
		C1	**X1**	**C2**	**X2**
Capital letters for emphasis	GREAT	0	9	2	13*
Word+space+dash+space+word	not – there	17*	5	37*	10
Two or more questions in row		4	6	5	14
Brackets used to show an afterthought or additional information	2 guys (i'd call them my angels)	2	2	4	3
Features not shared with Suspect X's text	**Type and examples**				
Punctuation, typography and spelling	Word+no space+dash+space e.g. not- there Word+no space+dash+no space+word e.g. friend-but Apostrophe omissions e.g. doesnt, dont, cant (C2/15*) Typographical errors e.g. mayb, willl, tought, cioa, pissd, w.end, I.ve, sh!t, Word initial lowercase e.g. **d**ecember, **s**yria, **s**unday, **e**aster				
Lexis	Colloquialisms e.g. wanna, outta, fab, fam				
Digitally mediated communication features	Use of lowercase *i* for first person singular e.g. And **i** want to hear (C1/20*)(C2/48*) Emoticons e.g. ☺ , :) , :(, ;) (C1/18*)(C2/34*) Other DMC features e.g. r (are), gr8, ya, yaself				

The second part of the table works with four categories that were identified as appropriate: *punctuation, typography and spelling*, *lexis* (including colloquialisms and idiosyncrasies)*, digitally mediated communication features* e.g. *u, cul8.* These categories highlight stylistic features which are not shared with the writer of the disputed text X. For reference purposes, the frequency counts are placed in brackets after the example, and an asterisk if there is a statistically significant difference.

The second part of the stylistic analysis discusses the results for each text and also, where relevant, briefly considers other aspects such greetings, salutations and

paragraphing, such as whether the writer consistently uses one or two-sentence paragraphs, or has longer paragraphs depending on the context of the discussion.

A feature which has not been considered for the qualitative analysis is that of grammar. The grammar used by the texts that is not standard written forms relate to the informal, speech-like register of the genre, such as the omission of the first person pronouns (*Would love to hear …..*(X)), and the omission of the auxiliary verb (*What you been doing….*(D)). These forms are common across the writers and typical of language used on Facebook, and for that reason it was decided not to have a separate category for grammatical idiosyncrasies. However, such a category would have been considered if the texts had been drawn from second language speakers or less well educated mother-tongue speakers.

3.7.2 Descriptors for forensic document analysis

McMenamin (2002, 124) states that “conclusions regarding authorship are stated in terms of identification or exclusion on a five-, seven-, nine point continuum” scale which has been accepted by the American Society for Testing and Materials (ASTM), or alternatively, one could employ a similar scale developed by the Scientific Working Group for Forensic Document Examination (SWGDOC, cited in McMenamin 2002, 124) (Table 3.2) to present conclusions regarding resemblance between known and questioned documents. Table 3.2 breaks down the descriptors and presents the criteria needed to draw a conclusion. It is a nine point scale, where nine would result in a positive identification if all the criteria were met. Bands six, seven and eight also positively identify the questioned text, but allow for some variance, as would be expected when analysing texts.

Table 3.2 – Criteria for conclusions on authorship questions (SWADOC) (McMenamin 2002,126)

	RESEMBLANCE Questioned Vs Known	CRITERIA	CONSISTENCY Questioned Vs Questioned
9	**IDENTIFICATION (did write)**	1. Substantial significant similarities in the range of variation 2. No significant dissimilarities 3. No limitations present: non-occurrence of variables, dissimilarities, quantity of writing	**DEFINITE (one writer)**
8	**HIGHLY PROBABLE (did write)**	1. Substantial significant similarities in the range of variation 2. No significant dissimilarities 3. Limitations are present: non-occurrence of variables, dissimilarities, quantity of writing	**HIGHLY PROBABLE (one writer)**
7	**PROBABLE (did write)**	1. Some significant similarities in the range of variation 2. No significant dissimilarities 3. Limitations are present: non-occurrence of variables, dissimilarities, individualising characteristics, quantity of writing	**PROBABLE (one writer)**
6	**INDICATIONS (did write)**	1. Few significant similarities in the range of variation 2. No significant dissimilarities 3. Limitations may be present: non-occurrence of variables, dissimilarities, individualising characteristics, quantity of writing	**INDICATIONS (one writer)**
5	**NO CONCLUSION (Inconclusive)**	1. Insufficient significant similarities in the range of variation 2. Insufficient significant dissimilarities in range of variation 3. Limitations may be present: non-occurrence of variables, individualising characteristics, quantity of writing 4. There may be similarities and dissimilarities	**NO CONCLUSION (Inconclusive)**
4	**INDICATIONS (did not write)**	1. Few significant dissimilarities in the range of variation 2. Limitations may be present: non-occurrence of variables, individualising characteristics, quantity of writing 3. There may be similarities	**INDICATIONS (more than one writer)**
3	**PROBABLE (did not write)**	1. Some significant dissimilarities in the range of variation 2. Limitations may be present, associated with: non-occurrence of variables, individualising characteristics, quantity of writing 3. There may be similarities	**PROBABLE (more than one writer)**
2	**HIGHLY PROBABLE (did not write)**	1. Substantial significant dissimilarities in range of variation 2. Limitations are present: non-occurrence of variables, individualising characteristics, quantity of writing 3. There may be similarities	**HIGHLY PROBABLE (more than one writer)**
1	**ELIMINATION (did not write)**	1. Substantial significant dissimilarities in range of variation 2. No limitations present: individualising characteristics, quantity of writing 3. There may be non-occurring variables 4. There may be similarities	**DEFINITE (more than one writer)**

The assigning of a band five would indicate that it was not possible to reach a conclusion due to inconclusive or insufficient data. Bands two, three and four effectively

eliminate the suspected author due to too many inconsistencies, even though the 'suspect' authors share some features with the disputed text. Band one is a definite elimination. I will be drawing my conclusions regarding the qualitative research using this particular band scale.

Notwithstanding the fact that the SWGDOC is being used to analyse the qualitative aspects of the different authors, it makes frequent reference to *significant differences.* The SWGDOC scales do not explicitly state that this significance should be statistically determined, which results in a fair degree of subjectivity in the interpretation of the band scales. However, in my study, where relevant occurrences in the data are sufficient for statistical testing, a Chi-square test is conducted to give further weight to the qualitative interpretations. The numerical figures are not shown on the tables, but if there is a significant difference, an asterisk is placed beside the example as shown in Table 3.1.

3.8 Quantitative analysis

This section will look at how the quantitative analysis was conducted, starting with a brief discussion of some of the problems relating to using quantitative analysis in authorship attribution. This is followed by a description of the use of WordSmith Tools (*WST),* and concludes with an overview of the Chi-square (χ^2) test.

3.8.1 Importance of the quantitative analysis

McMenamin (2002, 137) describes the statistical measurement of linguistic features as a "powerful complement" to the qualitative analysis, as a "quantitative analysis of their respective frequencies would provide the analyst with the mathematical tools needed to test whether such differences are significant, i.e. have less than a 5% (or even 1%) chance of having occurred randomly". However, just as qualitative analysis has its critics, so, too, does quantitative analysis. One objection to quantitative analysis is that some language features are difficult to identify as discrete, countable units; for example, in cases of deliberate obfuscation, where an individual may be deliberately attempting to copy another's writing habits, and in some cases may have little identifying weight. Despite that, a quantitative analysis has two definite advantages, as described by

McMenamin (2002, 138): “it will actually make decision making related to hypothesis testing easier and more precise, and it meets internal (methodological) as well as external (judicial) requirements for scientific evidence”.

3.8.2 WordSmith Tools

Guillén Nieto, et al (2008) state that the application of technology to analyse questioned documents has greatly helped the work of the forensic linguist within a legal setting by increasing the scientific reliability of qualitative data analysis with quantifiable data. Guillén Nieto et al. (2008) discuss software for forensic authorship identification as being divided into two categories: (1) software for detecting plagiarism and historical authorship investigation, exemplified by *JVocalyse, CopyCatch Gold, Signature Stylometric System, and* (2) software for general purpose text analysis, which includes *WST, Simple Concordance Program* and *Textanz.*

As my research will be using *WST*, I will be focussing solely on it. *WST* is an organic integrated suite of programs for examining the manner in which grammatical and lexical features act in a text. It was developed by Mike Scott of the University of Liverpool (GuillénNieto et al. 2008) and is currently (July 2012) on version 6.0. *WST* has at its core three main tools, namely: *Concord, KeyWord* and *Wordlist* and 12 utilities. The three core tools all have the facilities for analysing texts and obtaining statistical support and are discussed briefly in the sections which follow.

3.8.2.1 Keywords

Scott (2012, 177) states that “key-words provide a useful way to characterise a text or a genre”. The purpose of the *KeyWords* function is to compare two word lists: a *reference corpus* and a *study corpus*. The comparison results in a list of keywords where the frequencies are significantly different between the *reference corpus* and the *study corpus.* For example, if the word *the* has a frequency of 6% in the *reference corpus*, and 5% in the *study corpus* it will not be key, even though it may be the most frequently occurring word (Scott 2012). Within a forensic linguistic context, the *KeyWord* tool is useful for observing lexical similarities or differences between two texts and lastly,

noticing the writer's stylistic preferences (Guillén Nieto et al. 2008). Figure 3.1 is a screenshot of the KeyWord function, which shows firstly the keywords and their frequency counts of the study corpus, which in this case is one of the known texts (in this case Writer A). The column headed RC refers to the reference corpus, which is the disputed text. The disputed text was chosen as the reference corpus, as each of the other texts will be, individually, referenced against it. Traditionally, in authorship attribution, the reference corpus is usually larger and consists of the 'given' material, while the study corpus is the 'new' material, but in this simulation of a forensic situation, where all texts were the same size, this issue was not seen as being of major importance. The RC column shows the frequency counts of the keywords for that corpus. In Figure 3.1 the first keyword is the personal pronoun *I* and it occurs 93 times in the study corpus and only 29 times in the reference corpus (disputed text). Such a discrepancy means it is very key, and that results in a keyness value of 35.43. The higher the keyness value, the more key the word is. The values shown in red are considered negatively key because they identify frequency counts of keywords in the reference corpus that are significantly higher than in the study corpus.

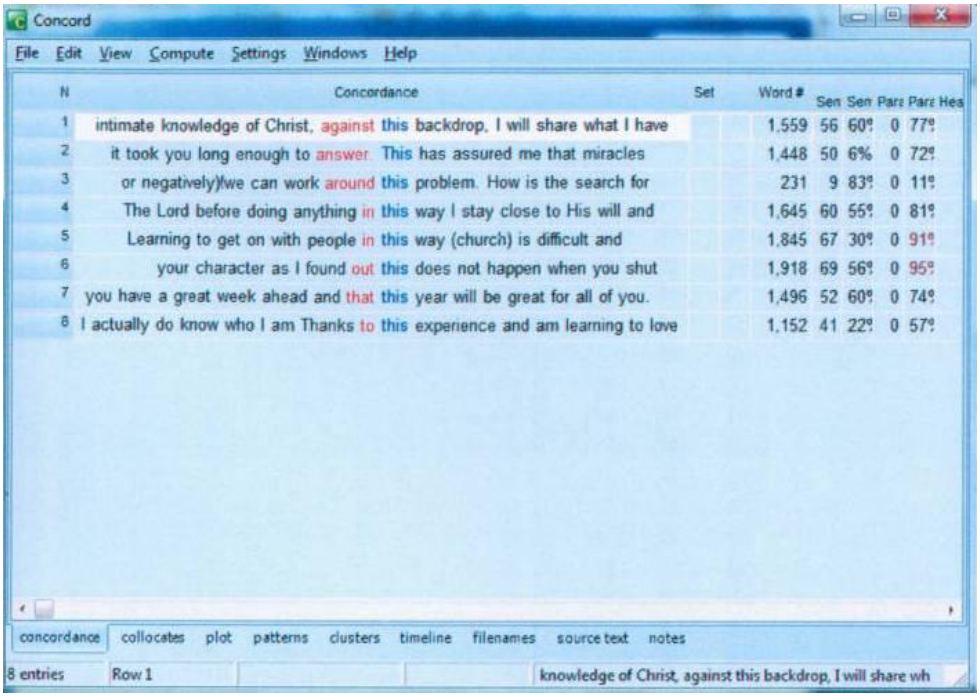
Figure 3.1 Screenshot from the* KeyWord *tool

N	Key word	Freq.	%	RC.	RC. %		P	Lemmas	Set
1	I	93	4.71	29	1.43	35.43	0.0000000003		
2	AM	21	1.06	1	0.05	17.05	0.0000363897		
3	WILL	25	1.27	3	0.15	16.47	0.0000493928		
4	NOT	29	1.47	5	0.25	16.36	0.0000523273		
5	WAS	16	0.81	3	0.15	7.96	0.0047733597		
6	GOD	8	0.40	0		6.34	0.0118071530		
7	MY	23	1.16	9	0.44	5.70	0.0169257000		
8	LORD	7	0.35	0		5.33	0.0210184567		
9	CHURCH	6	0.30	0		4.32	0.0377246775		
10	OUT	10	0.51	2	0.10	4.30	0.0381450057		
11	THAT	25	1.27	12	0.59	4.28	0.0385413207		
12	ARE	10	0.51	25	1.23	-5.26	0.0218723528		
13	LIKE	4	0.20	20	0.98	-9.01	0.0026864512		
14	AT	4	0.20	28	1.38	-16.01	0.0000628486		

3.8.2.2 Concord

Scott (2012, 124) states that “you get a much better idea of the use of a word by seeing lots of examples of it, and it’s by seeing or hearing new words in context lots of times that you come to grasp the meaning of most words in your native language”. The primary purpose of a concordance is to see many lexical items in context (Scott 2012). Despite the fact that this tool was not specifically designed for a forensic purpose, it can be used for forensic purposes as it allows a researcher to analyse a word, part of a word or lexical chunk in its linguistic context, and thereby notice recurring lexical features or idiosyncratic usages of a lexical item or chunk (Guillén Nieto et al. 2008). Figure 3.2 shows a screenshot of the *concordance* function for one of the writers with regard to *this*.

***Figure 3.2 – Screenshot of* Concord**



3.8.2.3 Wordlist

Scott (2012, 202) describes the purpose of the *WordList* tool as to (1) analyse the vocabulary used, (2) identify common word clusters, (3) compare the frequency of a word in different text files or across genres, (4) compare the frequencies of cognate

words or translation equivalents between different languages and, (5) lastly to get a concordance of one or more words in a list. *WordList* has a number of features useful to a forensic linguistic study. Guillén-Nieto, et al (2008: 16) explain a number of the useful functions of *WordList.* Firstly, it generates word listing in alphabetical order and/or frequency order, so texts can be analysed at a lexical level. Figure 3.3 is a screenshot of a wordlist from one of the participant writers and it exemplifies frequency of use option.

Figure 3.3 Screen of* WordList *showing the frequency list

3.8.3 Statistical methods and the Chi-square test

Grant (2007, 2) asks the question "how 'scientific' authorship analysis can and should be", particularly since the scientific aspect has, especially in the United States, been equated with quantifying both the analysis and the presentation of the results. In an authorship attribution context, quantification refers to the identification and frequency counting of selected linguistic features (style markers), which are then statistically measured in order to determine the origin of a text (Grant 2007). McMenamin (2002, 138) states that statistical tests are useful in "evaluating the significance of the relationship of variables across comparison writings", and the Chi-square test is

particularly useful for analysing the possible relationship between variables when rendered as frequency counts. Moreover, statistically measuring frequency counts would provide the researcher with the evidence of whether the differences are significant, i.e. whether there is less than a 5% or 1% chance of a specific feature or group of features occurring randomly (McMenamin 2002).

McMenamin (2002, 147) notes that the Chi-square test is used to evaluate the relative homogeneity of multiple variables expressed as actual frequencies in various questioned writings. This statistical test has been employed in numerous forensic linguistic cases, for example: Svartvik (1968), Hubbard (1995), McMenamin (2002) and Chaski (2001 and 2005). The Chi-square test tests the independence of two or more groups of frequency counts where there may or may not be a normal distribution (Chaski 2001). When utilising Chi-square, the size of the observed frequencies has to be considered as they are used to calculate the expected frequencies (Chaski 2001). If the total for the observed frequencies of an item in two texts is less than ten, then that particular item cannot be used (Cochran 1954). As Chi-square tests for difference between two sets of frequencies, the null hypothesis stipulates that there is no difference between the sets of frequency counts, and therefore the hypothesis of sameness can be accepted if the probability associated with Chi-square is greater than 0.05. The null hypothesis is rejected and it is accepted that there is difference if the probability is less than or equal to 0.05 (Chaski 2001). For the keyness Chi-square test calculations in my study, I use the Chi-square test calculations performed automatically by *WST* for the keyness analysis.

The statistical program *Instat,* which was developed by the *Statistics Service Centre* of The University of Reading in the United Kingdom, is used for the Chi-square test calculations needed for the function words, most frequently occurring words and punctuation. Table 3.3 shows an example of the Chi-square test calculations conducted by *Instat* (Instat Plus 2005) with regard to the disputed text (X) and Writer A. The shaded blocks indicate the function words which were omitted from the calculations, as their frequency counts did not add up to 10. It can be seen that seven of the 23 function

words were omitted, which underlines the potential difficulties in applying statistical tests to short texts.

Table 3.3 - Chi-square test grid for function words between the disputed text and Writer A at the 1,000-word level

Function word	Writer X observed	Writer X expected	Chi² value	Writer A observed	Writer A expected	Chi² value
A	21	15.9	1.636	13	18.1	1.437
All	2			7		
Also	0			3		
And	36	29	1.690	27	33.5	1.261
Any	0			3		
As	6	4.7	0.360	4	5.3	0.319
At	16	8.9	5.660	3	10.1	4.991
Been	7			1		
But	5	6.1	0.198	8	6.9	0.175
For	10	10.3	0.009	12	11.7	0.008
In	15	14.1	0.590	15	15.9	0.051
It	9	5.6	2.064	3	6.4	1.806
No	5	4.7	0.019	5	5.3	0.017
Not	1	8.0	6.125	16	9.0	5.444
Of	13	11.3	0.256	11	12.7	0.228
On	2	5.2	1.969	9	5.8	1.766
That	7	9.8	1.229	14	11.2	0.700
The	26	32.3	0.117	43	36.7	1.081
This	6			1		
To	29	30.9	0.117	37	35.1	0.103
Very	4			3		
Was	2	5.6	2.314	10	6.4	2.025
Were	1			2		
Chi Square value: 44.74						
Degrees of freedom: 15						
Significance level: 0.001						

3.9 Stylometric assessment of the participants' writings

This section deals with the practical quantitative research. My first task was to divide the 2,000-word submissions into two subsets: a 1,000-word document and a 2,000-word document, which I then had to save as text (.txt) documents to be accepted by *WST.* Appendix 3 contains the1,000-word and 2,000-word submissions. The 1,000-word cut-

off is clearly marked with a solid line. It should be noted that there is, on occasion, a slight discrepancy in the word count between MS Word and WordSmith Tools.

The first test to be conducted was the keyword test, the second test involved Morton's (1978) list of 23 function words, the third test performed was for the most frequently occurring words and the last test was an analysis of the punctuation. All four tests were conducted at both the 1,000-word level and 2,000-word level.

3.9.1 Test 1 – Keywords

The first test conducted was that of keyness. The *KeyWord* function of *WST* enables one to compare significantly different frequencies of lexical items across two texts. To achieve this, one text is designated the reference corpus, which for this study is the disputed text, and the other is the study corpus, which for this study is one of the known texts for Writers A to H. The column headed keyness gives the keyness value for each word, and the higher the keyness value, the more key (significantly different) a word is. The *KeyWord* function allows one to set the significance value (in my study $p \leq 0.05$). A word is considered positively key if it occurs significantly more frequently in comparison to the reference corpus, and is negatively key if it occurs significantly less frequently in comparison to the reference corpus. Negative keywords are highlighted in red (Scott 2012), as seen in Figure 3.4

Figure 3.4 Example of keyword test

N	Key word	Freq.	%	RC.	RC. %	Keyness	P
1	I	45	4.63	16	1.58	14.46	0.0001434341
2	NOT	17	1.75	1	0.10	13.24	0.0002743881
3	AM	8	0.82	0		6.44	0.0111636408
4	WAS	10	1.03	2	0.20	4.40	0.0359631218
5	WILL	8	0.82	1	0.10	4.27	0.0388669111
6	HAD	8	0.82	1	0.10	4.27	0.0388669111
7	MY	13	1.34	4	0.40	4.13	0.0420979820
8	AT	3	0.31	16	1.58	-7.17	0.0073951408

3.9.2 Test 2 - Function words

The second test involves an analysis of 23 function words as postulated by Morton (1978), namely: *a, all, also, and, any, as, at, been, but, for, in, it, no, not, of, on, that, the, this, to, very, was* and *were*. Morton selected these 23 function words as he believed them to be the most commonly used ones. Moreover, these function words were also chosen because they are believed to be used subconsciously and are therefore not generally under the overt control of the writer (as discussed in Chapter 2). It is worth noting that these function words were drawn up for stylometric analysis of novelists (Hubbard 1995) rather than the DMC of Facebook, which is a very different genre. Despite Morton having been discredited because of his CUSUM technique, his list of function words has been used by others (Hubbard 1995), and in my study the list was included to test its relative usefulness against the other tests. In order to obtain the frequency counts of these 23 function words, I had to enter each writer's submission individually into the *WordList function* of *WST* to create a word list for each individual writer, *a*s exemplified in Figure 3.5.

Figure 3.5 Example of a wordlist

File Edit View Compute Settings Windows Help

N	Word	Freq.	%	Texts	%	Lemmas
1	I	45	4.63	1	100.00	
2	THE	43	4.42	1	100.00	
3	TO	38	3.91	1	100.00	
4	AND	27	2.78	1	100.00	
5	YOU	19	1.95	1	100.00	
6	NOT	17	1.75	1	100.00	
7	IN	15	1.54	1	100.00	
8	IS	14	1.44	1	100.00	
9	THAT	14	1.44	1	100.00	
10	A	13	1.34	1	100.00	
11	MY	13	1.34	1	100.00	
12	FOR	12	1.23	1	100.00	
13	BE	11	1.13	1	100.00	
14	OF	11	1.13	1	100.00	

Using the frequency counts obtained from the wordlists, I was able to populate the tables (Table 3.4 shows an abridged example). The second row shows all the function words from the list. The frequency counts for each writer's (A-H) use of that function word are placed adjacent to the function word. For example, Writer X had 21 examples of the indefinite article *a* to Writer A's thirteen examples. Where there is a significant difference of p≤ 0.05 the block is shaded green and where there is a very significant difference of p≤ 0.01 the block is shaded red. This bird's-eye view gives us a "visual impression of the degree of difference between each text and the X corpus as well as an indication of where the individual differences are" (Hubbard 1995, 60). Both corpora were subjected to a Chi-square test, where the profile of the disputed text (X) was compared to each of the other writers' profiles.

Table 3.4 Extract from the table used to analyse function words

		X	A	B	C	D	E	F	G	H
1	**A**	**21**	13	29	24	29	16	14	23	10
2	**All**	**2**	7	2	7	0	5	8	5	8
3	**Also**	**0**	3	1	3	1	3	6	5	2
4	**And**	**36**	27	35	16	25	34	23	24	12

3.9.3 Test 3 - Most frequently occurring words

The third test is an analysis of the most frequently occurring words and is, in certain respects, an extension of function words and Kotzé's (2010) work with keywords. Unlike a function word list, this test draws its lexical items directly from the texts being analysed rather than having a predetermined list. The test identified the 30 most frequently occurring words in the disputed text (X), and compared them to the frequencies of those same words in the other eight texts and put them into tabular form, as shown in Table 3.5.

Table 3.5 – Extract from the most frequently occurring words table

		X	A	B	C	D	E	F	G	H
1	**A**	**36**	39	55	41	50	42	36	39	28
2	**About**	**19**	10	3	15	10	6	8	12	7

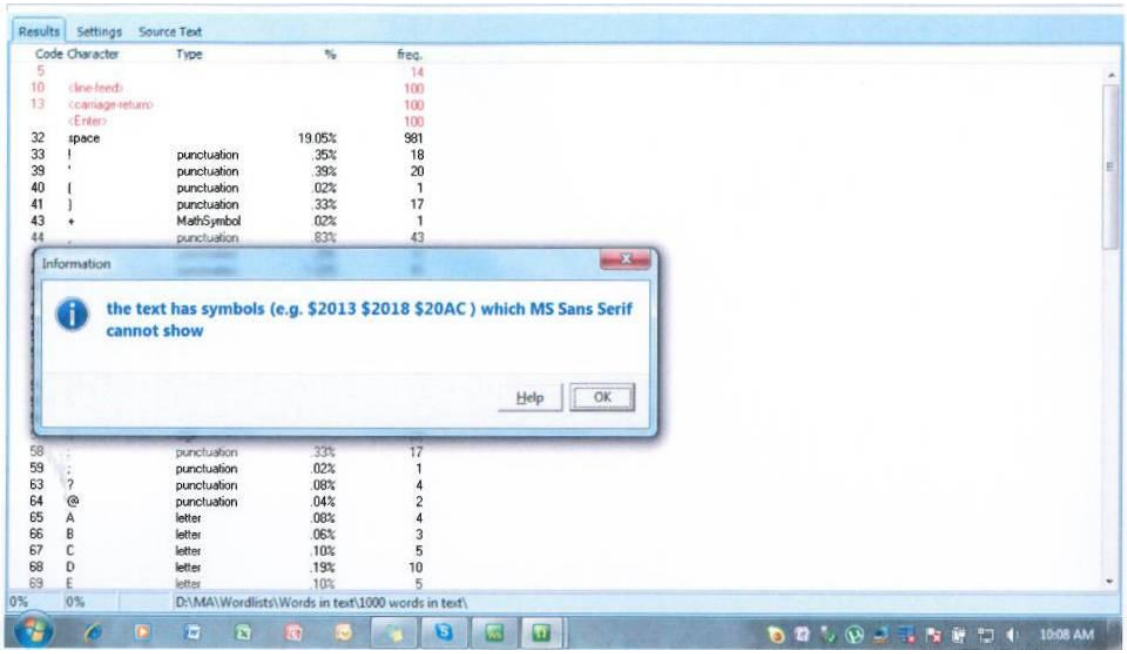
The blue shaded areas are Morton's (1978) function words, as some of them, as would be expected, occurred in the most frequently occurring words and have been highlighted for reference purposes only. The results from the keyword analysis were used to determine which of the most frequently occurring words were key. Where there is a significant difference of p≤ 0.05, the block is shaded green, and where there is a very significant difference of p≤ 0.01, the block is shaded red. Both the1,000-word and 2,000-word corpora were subjected to a Chi-square test using the statistics program *InStat*. The profile of disputed text (X) is compared to each of the other writers' profiles for these words and the overall results are placed at the bottom of the table below the corresponding author.

3.9.4 Test 4 – Punctuation

The fourth test involves analysing punctuation, which was informed by research conducted by Chaski (2005) and Olsson (2008). As noticed in the stylistic assessment, the writers in my study exhibit numerous non-standard punctuation styles, with multiple question marks, exclamation marks, ellipsis and question mark/exclamation mark combinations being the norm, and in addition to that, there were several different emoticons (☺, :-), :-P, etc) on display. All of these non-standard features rendered a normal frequency count problematic. Therefore, I categorised the features as sets, for example, if four question marks combined marked the boundary of a sentence, then those four question marks were treated as a single unit. For this analysis I was unable to use *WST,* because although it is true that *WST* can do frequency counts of punctuation in the *Character Profiler* utility, it only counts the individual frequency of each feature and not sets. The *Concordance Tool* was only marginally better in that it could give examples of most punctuation marks, but could not analyse question marks as the program considers the question mark to represent an unknown and gives examples of all words in alphabetical order. Moreover, when converting a *Microsoft Word* document to a *plain text* document certain features may be lost, e.g. Writer E had the following sentence: *lots of love* ☺ ☺ ☺, which was rendered in plain text as *lots of love* :-) :-) :-), and there were symbols which could not be shown. Figure 3.6 shows a

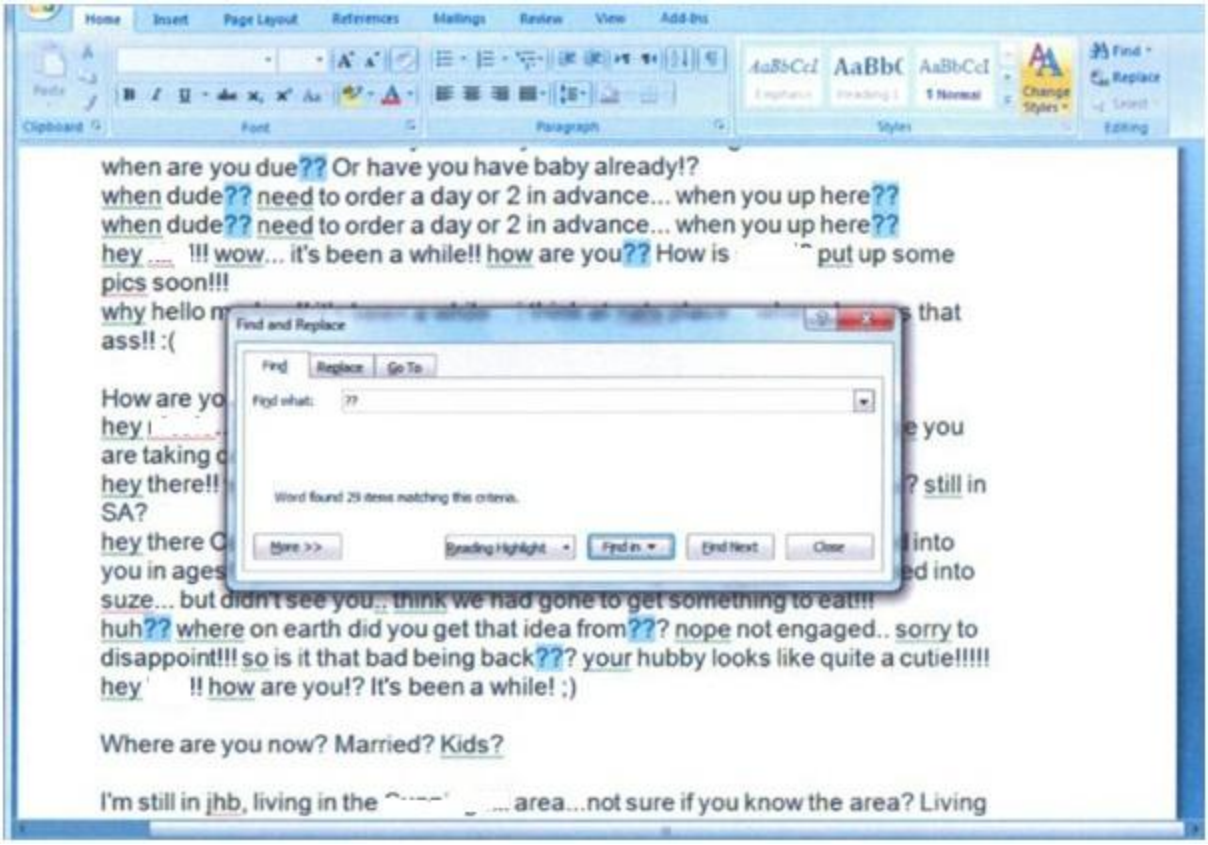
screen shot from the *Character Profiler utility* showing *WST's* limitation in analysing punctuation.

Figure 3.6 – Character Profiler utility



Due to the fact *WST* was not able to analyse punctuation in the manner my study required, I used the search function of *Microsoft Word.* This function has the advantage of being able to display all of the features from the original document as they are highlighted in the text and then counted manually. Unfortunately, this process has the potential to be problematic, as the search function will show all examples, even where they may be connected to something else. For example, if one searches for double question marks, then it will highlight all double question marks even if there is a third question mark (???), as can be seen in Figure 3.7. In this example, I would not count this as an example of a double question mark, but of course rather as a triple question mark feature.

Figure 3.7 – Microsoft word search function



Once all the punctuation features had been counted they were put into a table. There is one table for the1,000-word level and another table for the 2,000-word level. The data in both tables were subjected to a Chi-square test, where the profile of the disputed text (X) is compared to each of the other writers' profiles. Any frequency counts where the combined disputed and known feature were below 10 (meaning that they did not meet the Chi-square requirement of an expected frequency of at least 5), were omitted from the analysis. The Chi-square results of each writer's profile were calculated using the *InStat* program and placed at the bottom of the table beneath the relevant author. The features marked with an asterisk are a set and are considered as a single unit. Table 3.6 shows how the punctuation analysis is displayed.

Table 3.6 Extract of the table used to analyse punctuation

Feature	e.g.	X	A	B	C	D	E	F	G	H
Full stop	.	**61**	29	30	38	17	26	61	58	33
Comma	,	**15**	9	23	17	27	9	43	15	49
Single quotes*	'….'	**4**	4	1	0	1	0	1	0	1

3.10 Conclusion

The aim of this chapter was to present the research methodology employed to conduct the study. It began by examining the ethical dilemmas emanating from the study. As I am using real texts written for genuine social interaction, rather than fabricated ones for testing purposes, I ensured that all aspects of my participants' privacy were maintained by avoiding all names, telephone numbers, e-mail and Skype addresses. My participants were chosen according to stringently defined criteria: they were all female, between 30 and 40 years, spoke English as a home language, were educated to post-matric level and were active users of Facebook.

The aim of this study is to explore the extent to which it is possible to attribute authorship on a social networking site, in particular Facebook. To achieve that aim, I needed genuine text written on Facebook. Each of my eight participant writers submitted 2,000 words of text. As this is a simulated case, I needed one participant to take on the role of the accused and to submit an extra 2,000 words, which was designated disputed text X, against which the other texts would be compared. It may have been better if I had asked someone else to select the participant who was to provide the disputed text, so as to avoid bias on my part. Unfortunately, this would have been somewhat difficult from a practical perspective, as they do not all know each other. Having said that, though, any bias on my part should not have any bearing on the quantitative assessment, and even the qualitative assessment consisted of features that were subjected to frequency counts. I did not alter the content or syntax of the texts apart from changing any names and removing any contact information. The texts were then divided into two groups, consisting of the first 1,000 words and the whole 2,000 words respectively.

Authorship attribution within the field of forensic linguistics involves comparing a disputed text against a known authored text or a series of texts. The texts can be analysed either stylistically or stylometrically, or both methods may be used. My research, being a simulation to test the efficacy of certain style markers, made use of both. Starting with the qualitative analysis, the chapter looked at a number of stylistic features such as punctuation, typography, spelling, lexis, forms of digitally mediated communication and grammar. The quantitative approach to authorship attribution is also not without its drawbacks. Despite the benefit of numerically evaluating stylometric features, which makes hypothesis testing simpler and more accurate (McMenamin 2002), there is often the problem in forensic cases of the text size being too small to statistically evaluate. One criticism of the stylistic approach is that it lacks scientific rigour and is based on ill-defined interpretation (Chaski 2005). To counter this to an extent, I have introduced the Scientific Working Group for Forensic Document Examination (SWGDOC) band scales, which give fairly clear guidelines on what is expected before a certain band score can be awarded.

Stylometric analyses of texts have become considerably easier in recent years with the advent of computer programs designed to aid in the analysis of texts. It has been explained in this chapter that this study made use of the three functions of WordSmith Tools (WST) namely: *Keywords, Concordance* and *WordList*. The *Keyword Tool* was used to find keywords which were significantly more frequent between the disputed text (X) and each of the texts A-H in turn. The *Concordance Tool* was used to see words in their contexts, and the *WordList Tool* was used to count the frequencies of words as they occur in the texts. The stylometric analysis of the texts started with keywords, followed by Morton's (1978) function words, then the most frequently occurring words and lastly punctuation.

The following chapter will show how the research methods described here were applied and will present the findings of the study.

Chapter 4 – Findings

> *While it is possibly true that mistakes made by authorship analysts in the field of literature could lead to red faces and bad press at worst, the same cannot be said of the forensic context, where mistakes could lead to imprisonment or even execution in some countries. The importance of extreme caution before arriving at conclusions can therefore not be overemphasised.*
>
> (Kotzé 2010, 186)

4.1 Introduction

This chapter presents the findings of the study, covering first those for the qualitative analyses and then those for the quantitative analyses before moving on to conclusions.

4.2 Qualitative analysis

This section looks at the findings from the qualitative analysis. Each writer's submission was judged against the disputed text, and then appraised, using the SWGDOC band scale. X's profile will be presented first (Table 4.1), highlighting the distinctive features of her writing, followed by the profiles of Writers A to H, which are compared to X. For each analysis, I begin firstly by describing the similarities that are noticed, and secondly the differences, and then weigh them up to reach a conclusion. The comparisons are done at both the1,000-word (e.g. X1) and 2,000-word (e.g. X2) levels. Even though the approach here is essentially qualitative, an asterisk has been placed beside any counts which meet the Chi-square requirement of an expected value of at least 5 and also exhibit statistical significance.

4.2.1 Writer X

Table 4.1 – Writer X: highlighted features

Features	Examples	Frequency X1	X2
Capital letters for emphasis	SOOOOOO, WOW	9	14
Words/phrases highlighted in single quotes	'*name*'	4	6
Word+space+dash+space+word	level – don't	5	10
Word+no space+ellipsis (4 dots)+no space+word	anything....really	0	5
Two or more questions in a row		6	14
Brackets used to show an afterthought or give additional information	(like 2 year olds a lot of the time)	2	3
Apostrophe + cause	'cause	0	1
Other punctuation, typography and spelling	Spelling error/mistake Lead for led		
Lexis	Cangle (cankle) cankle - noun; the meeting of the calf and the foot where an ankle is not present due to lack of ankle definition (urbandictionary.com)		
Digitally mediated communication features			

Writer X exhibited excellent control over grammar and lexis, and although the style was chatty and informal, standard conventions were generally maintained. However, there were a few noticeable features, the most noticeable being X's large number of questions, usually several in a row. She also chose to show emphasis by way of using upper-case letters and in some instances extending the word (SOOOO). She tended to follow the standard letter layout for all her posts with a greeting which was always *hi* + recipient's name followed by several body paragraphs and a salutation, which was either *Lots of love, Take care* or *Love*. Turning to the themes of the postings: the first six postings by X focussed exclusively on her new job in a new country, whereas, in the final four postings, she described her anxieties about moving and was even asking for advice, and it was here that we find the bulk of the question forms.

4.2.2 Writer A/Writer X

Table 4.2 - Writer A/Writer X

Features shared with Suspect X's text	**Examples**	**Frequency**			
		A1	**X1**	**A2**	**X2**
Capital letters for emphasis	INFLUENCING	0	9	3	14*
Words/phrases highlighted in single quotes	'sucked into FB'	4	4	4	6
Word+space+dash+space+word	Tickets – but he has – so	43*	5	52*	10
Word+no space+ellipsis (4 dots)+no space+word	dry....I	1	0	3	5
Two or more questions in row		2	6	3	14*
Brackets used to show an afterthought or give additional information	(my opinion)	10	2	16*	3
Features not shared with suspect X's text	**Type and examples**				
Punctuation, typography and spelling	Words/phrases highlighted in double quotes e.g. "Sheep on drugs" Word+no space+ellipsis+space (definite end of sentence) e.g. No thanks..... Incorrect use of apostrophe e.g. loo's (plural) Non-standard word-initial capitalisation e.g. **E**ncouraging (A1/22*)(A2/79*)				
Lexis					
Digitally mediated communication features					

Writer A shared five features with X at the1,000-word level and six feature at the 2,000-word level, although some of these features occurred significantly less or more often than in X. She used the construction: *word+space+dash+space+word* just over eight times more at the 1,000-word level, and five times more than X at the 2,000-word level. Writer A used *brackets to show an afterthought or give additional information* five times as often at both the 1,000-word and 2,000-word levels as X. Writer A used the feature *capital letters for emphasis*, but significantly less than X at the 2,000-word level, and she did not exhibit any examples of this feature at the1,000-word level.

Despite the similarities, there were a number of features used by A which were not shared by X and occurred in large enough frequencies to be significant (the figures with asterisks and in brackets represent the frequencies that differ significantly from X at the 1,000-word and 2,000-word levels respectively). The most noticeable of A's dissimilarities from X was her use of word-initial capital letters (***R**eunion*). Notwithstanding the fact that a large portion of her writing was devoted to religious topics, where it is common to use word-initial capitalisation when describing God, Jesus. etc., there were still many examples where such a construction would be considered marked (*Take Care*). Writer A also ended a number of sentences with a series of dots rather than a single full stop or other standard boundary marker. Finally, she had at least seven examples of incorrect apostrophe usage, usually to denote plurality (*loo'**s***), and apostrophes were missing where they should have been present (*dont*).

I would score A as a Band 2 on the SWGDOC scale for both the1,000-word and 2,000-word levels, which means that it is highly probable that A is not X, as the text meets the criteria of substantial significant dissimilarities in range and variation, even though there were some similarities. Moreover, there were significant individualising characteristics, particularly regarding apostrophe usage and word-initial capitalisation to warrant A being rejected as X.

4.2.3 Writer B/Writer X

Table 4.3 - Writer B/Writer X

Features shared with Suspect X's text	**Examples**	**Frequency**			
		B1	**X1**	**B2**	**X2**
Word+space+dash+space+word	Okay confirmed – 20th	3	5	6	10
Two or more questions in a row		8	6	14	14
Brackets used to show an afterthought or give additional information	(excuse the analogy)	3	2	4	3
Features not shared with Suspect X's text	**Type and examples**				
Punctuation, typography and spelling	Word+no space+ellipsis (End of sentence/paragraph) e.g. for a picnic... (B1/17*)(B2/25*) Word+no space+ellipsis+smiley e.g. fasting....:) General spelling errors e.g. tomatoe, ridiculus, shakey Word initial lowercase e.g. **s**unday, **m**onday, **r**amadan				
Lexis	Idiosyncrasies e.g. anyhoo, insy winsy tini tiny bitsy witsy Colloquialisms e.g. nite, gonna, wanna, kinda, gotta				
Digitally mediated communication features	Pls, ya				

Writer B shared four features with X at both the1,000-word and 2,000-word levels. What is more, none of the shared features had frequency counts high enough to be considered significant. Despite these similarities, however, there were some substantial differences, the most noticeable being: B's 17 uses of the construction *word+no space+ellipsis* to mark the boundary of a sentence or paragraph at the1,000-word level and 25 at the 2,000-word level, and the seven instances where the boundary marking ellipsis was concluded with an emoticon. Moreover, Writer B exhibited many of the features generally associated with DMC such as *pls* (please), *ya* (you) as well as numerous colloquialisms such as *gonna, wanna* and *kinda*. There were also a number of spelling errors (*tomatoe, ridiculus)* and words which should begin with uppercase letters and did not, such as *sunday.* Lastly, she had a number of examples of present participles without the final 'g', such as *stayin* and *crashin.* Writer B extended her chatty

style of writing to include features which describe paralinguistic activities that supposedly accompany her writing (*heeheehee* and *sniff sniff sniff).*

I would score Writer B as a Band 2 for both the1,000-word and 2,000-word levels on the SWGDOC scale, which means it is highly probable that Writer B is not Writer X as the text meets the criteria of substantial significant dissimilarities in range and variation even though there were some similarities. Moreover, there were significant individualising characteristics, particularly regarding spelling errors, colloquialisms and punctuation.

4.2.4 Writer C/Writer X

Table 4.4 - Writer C/Writer X

Features shared with Suspect X's text	**Examples**	**Frequency**			
		C1	**X1**	**C2**	**X2**
Capital letters for emphasis	GREAT	0	9	2	14*
Word+space+dash+space+word	not – there	17*	5	37*	10
Two or more questions in row		4	6	5	14
Brackets used to show an afterthought or additional information	2 guys (i'd call them my angels)	2	2	4	3
Features not shared with Suspect X's text	**Type and examples**				
Punctuation, typography and spelling	Word+no space+dash+space e.g. not- keep Word+no space+dash+no space +word e.g. friend-but Apostrophe omissions e.g. doesnt, dont, cant (C2/15*) Typographical errors e.g. willl, tought, cioa, pissd, w.end, I.ve, sh!t, Word initial lowercase e.g. **d**ecember, **s**yria, **s**unday, **e**aster				
Lexis	Colloquialisms e.g. Wanna, outta, fab, fam				
Digitally mediated communication features	Use of lowercase *i* for first person singular e.g. And **i** want to hear (C1/20*)(C2/48*) Emoticons e.g. ☺ , :) , :(, ;) (C1/18*)(C2/34*) Other DMC features e.g. r (are), gr8, ya, yaself				

Writer C shared three features with X at the1,000-word level and four at the 2,000-word level. However, she used the construction: *word+space+dash+space+word* almost three times as much at the1,000-word level and four times as much as X at the 2,000-word level. On the other hand, Writer X also used *capital letters for emphasis* just over six times as much as C at the 2,000-word level.

Despite those few similarities, Writer C exhibited numerous dissimilarities to writer X, with the most prominent being in her choice of lexis. Writer C made frequent use of features usually associated with DMC, such as her 48 examples of *i* for the first person

singular pronoun. Writer C also exhibited a number of idiosyncratic spellings and word forms (*w.end* and *should ve)*, and there were also 15 examples of incorrectly used or omitted apostrophes. She also preferred the constructions of *word+no space+dash+space* (*not- there*) and *word+no space+dash+no space+word* (*friend-but*). Lastly, Writer C had 34 examples of emoticons, which included: ☺ , :) , :(.

I would score Writer C as Band 1 (elimination) for both the1,000-word and 2,000-word levels on the SWGDOC scale, as the text meets the criteria of substantial significant dissimilarities in range and variation. Even though there were a few similarities, two out of the three similarities exhibited quite extreme significant differences. Moreover, there were significant individualising characteristics, particularly regarding numerous spelling idiosyncrasies, colloquialisms, emoticon use, and layout of text as well as very different ellipsis dot constructions.

4.2.5 Writer D/Writer X

Table 4.5 - Writer D/Writer X

Features shared with Suspect X's text	**Examples**	**Frequency**			
		D1	**X1**	**D2**	**X2**
Capital letters for emphasis	CONGRATS	4	9	7	14
Words/phrases highlighted in single quotes	'normal'	1	4	6	6
Two or more questions in a row		16	6	20	14
Features not shared with Suspect X's text	**Type and examples**				
Punctuation, typography and spelling	Word+no space+ellipsis+space (3 dots) e.g. wow... it's (D1/47*) (D2/96*) Word+no space+ellipsis+space (2 dots) e.g. doing well.. but (D1/31*)(D2/83*) Non standard end of sentence punctuation • exclamation mark+question mark (!?) (D1/16*)(D2/22*) • double question marks (??) (D1/19*)(D2/28*) • triple question marks (???) • double exclamation marks (!!)(D1/13*)(D2/36*) • Triple exclamation marks (!!!)(D1/16*)(D2/33*) Typographical error e.g. assuiming Word initial lowercase e.g. sunday, october, friday, november, london, chinese				
Lexis					
Digitally mediated communication features	Emoticons e.g. :) , :(, :P , :D Use of lowercase *i* for first person singular e.g. when i get back				

Writer D shared three features with Writer X at both the1,000-word and 2,000-word levels, and without any features being significantly more or less frequent.

Despite the similarities, Writer D exhibited numerous dissimilarities to Writer X, with some substantial frequency counts. The most noticeable are her use of constructions as sentence boundary markers: *word+no space+ellipsis* (3 dots)+*space* and *word+no*

space+ellipsis (2 dots)+*space*. Another very noticeable feature of Writer D was her use of non-standard boundary markers, particularly *exclamation mark+question mark*, *double question marks*, *triple question marks*, *double exclamation marks* and *triple exclamation marks*. Furthermore, Writer D habitually added various emoticons to the end of her sentences. She also used the lower case *i* when referring to the first person singular pronoun *I,* although this was not absolute, as she often used the standard form of *I.* A further consideration was her use of word-initial lowercase: *sunday*, when the standard calls for uppercase *Sunday*.

I would score Writer D as a Band 2 for both the1,000-word and 2,000-word levels on the SWGDOC scale, which means that it is highly probable that Writer D is not Writer X, as it meets the criteria of substantial significant dissimilarities in range and variation, even though there were some similarities. Moreover, there were significant individualising characteristics, particularly regarding her creative means of signalling sentence boundaries.

4.2.6 Writer E/Writer X

Table 4.6 - Writer E/Writer X

Features shared with Suspect X's text	**Examples**	**Frequency**			
		E1	**X1**	**E2**	**X2**
Capital letters for emphasis	SO	2	9	4	14*
Word+space+dash+space+word	day – she	20*	5	53*	10
Two or more questions in a row		11	6	20	14
Brackets used to show an afterthought or give additional information	(nearly 2 years here - whew)	12*	2	18*	3
Features not shared with Suspect X's text	**Type and examples**				
Punctuation, typography and spelling	Word+no space+ellipsis+space (3 dots) e.g. time goes... one minute (E1/22*)(E2/49*) Word+no space+ellipsis+space (4 dots) e.g. stories.... (End of paragraph) (E2/13*) Typographical errors e.g. bussiness, thw consdiered				
Lexis	Idiosyncrasies e.g. famdamily, cuz, parentals, favouritist Colloquialisms e.g. wif, vol kak				
Digitally mediated communication features	ya (you)				

Writer E shared four features with X at both the1,000-word and 2,000-word levels, but there were some significant frequency differences in their usages. Writer E's use of the construction: *word+space+dash+space+word* exhibited significant differences at both levels. At the 1,000-word level, E used it four times as often as X and just over five times as often as X at the 2,000-word level. Writer E also employed brackets to show an afterthought or to give additional information six times more than X and significantly more often at both levels. Writer X, on the other hand, used capital letters for emphasis considerably more often at both levels.

Despite the four shared features, there were a few marked features E employed, particularly regarding ellipsis dots. She made use of the following constructions either intrasententially, or as a boundary marker: *word+no space+ellipsis* (3 dots)+*space* and *word+no space+ellipsis* (4 dots)+*space* at frequencies of 49 and 13 respectively at the

2,000-word level. The first feature was statistically significant at both levels, and although the second feature was not significant enough at the1 000-word level, there were a fair number of instances of this feature even at this level. Writer E also employed some very idiosyncratic lexis (*famdamilies, favouritist*) and some colloquialisms (*wif, vol kak).* (It would not be unusual for the participants in this study to have used Afrikaans colloquialisms, as all of them would have studied Afrikaans at school. Furthermore, colloquial Afrikaans, e.g. *lekker* (nice), and African language, e.g. *fundi* (teacher/expert) lexis is common in South African English). Moreover, she also wrote out her paralinguistic activities (*heeheehee, achooooo*).

I would score E as a Band 2 at both the1,000-word and 2,000-word levels on the SWGDOC scale, which means it is highly probable that E is not X as it meets the criteria of substantial significant dissimilarities in range and variation even though there were some similarities. Moreover, there were significant individualising characteristics, particularly regarding her idiosyncratic lexis.

4.2.7 Writer F/Writer X

Table 4.7 - Writer F/Writer X

Features shared with Suspect X's text	**Examples**	**Frequency**			
		F1	**X1**	**F2**	**X2**
Capital letters for emphasis	DOES	2	9*	5	14
Word+space+dash+space+word	moved – me	9	5	23*	10
Two or more questions in a row		1	6	2	14*
Brackets used to show an afterthought or give additional information	(don't mean to be a bitch but he DOES)	1	2	2	3
Features not shared with Suspect X's text	**Type and examples**				
Punctuation, typography and spelling					
Lexis					
Digitally mediated communication features	Sentence initial emoticons e.g. :) , :-) , :-D , ☺ Intersentential emoticons e.g. :) , ☺ Sentence ending emoticons e.g. :) , ☺ (F2/16*) LOL (Laugh Out Loud)				

Writer F shared four features with X and only the construction: *word+space+dash +space+word,* which F used just over twice as much as X*,* and *capital letters for emphasis* show any significant frequency difference, with the latter only marginally significant at the1,000-word level.

Notwithstanding F's textual similarities to X, there were a number of dissimilarities that need further discussion. The most noticeable feature of F's texts was her use of emoticons, not just at the end of sentences, but at the beginning of sentences and intrasententially, which was juxtaposed with a register not as informal as many of the other writers (*but can also assist you with booking tickets and suchlike to other destinations in SA*). She made use of the construction *adjective+noun* (*stunning city, beautiful place*) more so than any of the other writers. Writer F had a combination of lengthy letter style texts as well as a number of very short one-or-two sentence texts.

Interestingly, it was in the shorter texts where one found more colloquial writing and the use of common DMC features such as *LOL*.

I would score Writer F as a Band 3 for both the1,000-word and 2,000-word levels on the SWGDOC scale, which means it is probable that Writer F is not Writer X, as the text meets the criteria of substantial significant dissimilarities in range and variation, even though there were some similarities. Moreover, there were significant individualising characteristics, particularly regarding her idiosyncratic use of emoticons. However, her consistent use of almost completely standard lexical forms brings F closer in style to X.

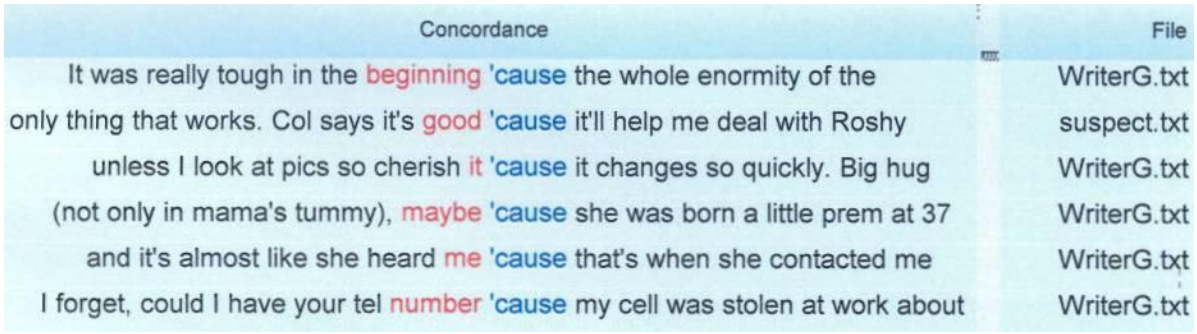
4.2.8 Writer G/Writer X

Table 4.8 - Writer G/Writer X

Features shared with Suspect X's text	**Examples**	**Frequency**			
		G1	**X1**	**G2**	**X2**
Capital letters for emphasis	EVER	1	9*	2	13*
Words/phrases highlighted in single quotes	'Phil'	0	4	1	6
Word+space+dash+space+word	Job – what	3	5	4	10
Word+no space+ellipsis (4 dots)+no space+word	You....the mother	8	0	11	5
Two or more questions in a row		6	6	13	14
Brackets used to show an afterthought or give additional information	(dealing with teething baby)	2	2	7	3
Apostrophe + cause	'cause	2	0	5	1
Features not shared with Suspect X's text	**Type and examples**				
Punctuation, typography and spelling	Word-initial lowercase e.g. **f**riday, **m**onday				
Lexis	Colloquialisms e.g. cos, aint, gonna				
Digitally mediated communication features					

Writer G shared six features with X at the1,000-word level and seven features at the 2,000-word level. The only feature to show any statistical significance is X's examples of using uppercase to show emphasis, which she used more often than G at both levels. Both authors had similar layout styles, greetings and salutations. Both authors followed the standard letter format and began each post with *Hi* and ended with either *Lots of love, Take care* or *Love.* It is true, though, that stylistic features such as layout and greetings can be easily copied and should be treated with caution. One feature that really stood out is the construction *apostrophe+cause* (*'cause*), which was used by G five times and X once. What makes this construction so marked is that according to the British National Corpus (Burnard 1995), it occurs only 1.27 times per million words. Figure 4.1 is a screen shot of *'cause* from the Concordance Tool of WordSmith Tools, which shows the examples as they appear in Writer G and Writer X's 'suspect' texts.

Figure 4.1 – Concordance for 'cause



Concordance	File
It was really tough in the beginning 'cause the whole enormity of the	WriterG.txt
only thing that works. Col says it's good 'cause it'll help me deal with Roshy	suspect.txt
unless I look at pics so cherish it 'cause it changes so quickly. Big hug	WriterG.txt
(not only in mama's tummy), maybe 'cause she was born a little prem at 37	WriterG.txt
and it's almost like she heard me 'cause that's when she contacted me	WriterG.txt
I forget, could I have your tel number 'cause my cell was stolen at work about	WriterG.txt

However, there were a few features exhibited in G that were not displayed in X. G had two examples of word-initial lowercase: ***m****onday* and ***f****riday*, where the standard calls for uppercase: ***M****onday* and ***F****riday.* In X, though, there are no mentions of days of the week, so this is not necessarily a difference between the two writers). Secondly, G had a number of examples of colloquialisms (***cos, ain't*** and *gonna*). It should be noted that the frequencies of the aforementioned features were extremely small.

I would score Writer G as a Band 7 on the SWGDOC scale for the1,000-word level, meaning that it is probable that Writer G is writer X. What prevents a Band 8 from being assigned are the differences in *Capital letters for emphasis.* Moreover, there were no shared examples of *words/phrases in single quotes* and the construction *word+no space+ellipsis (4 dots)+no space+word.* A Band 8 would be assigned to the 2,000-word level, meaning that it is highly probable that Writer G is Writer X as the text meets the criteria of substantial significant similarities in range and variation. What prevents a Band 9 from being awarded are the examples of words written in capital letters.

4.2.9 Writer H/Writer X

Table 4.9 - Writer H/Writer X

Features shared with Suspect X's text	**Examples**	**Frequency**			
		H1	**X1**	**H2**	**X2**
Capital letters for emphasis	ALMOST	1	9*	9	14*
Words/phrases highlighted in single quotes	'find a friend'	1	4	2	6
Two or more questions in a row		6	6	8	14
Brackets used to show an afterthought or give additional information	(everything's green)	5	2	10	3
Features not shared with Suspect X's text	**Type and examples**				
Punctuation, typography and spelling	Word+space+ellipsis (3 dots)+no space+word e.g. busy ...and (H2/29*) Word+space+ellipsis (4 dots)+no space+word e.g. e-mail ....like Typographical e.g. fortuntes, startnig General spelling errors e.g. ceilling, dolfins, hybernation Asterisks to show paralinguistic activities e.g. *hugs*				
Lexis	Idiosyncrasies e.g. Anyhoo Colloquialisms e.g. kinda				
Digitally mediated communication features	lol (Laugh Out Loud) Emoticons e.g. ;) , :D				

Writer H shared four features with Writer X, and all the frequencies except those for *capital letters for emphasis* at the 1,000-word level were too small to be significant. Despite these similarities, there were a number of textual features employed by Writer H not shared with Writer X. Only one feature stood out as particularly significant, which was the construction *word+space+ellipsis* (3 dots)+*no space+word*. Writer H also used two features commonly associated with DMC, namely: asterisks to show paralinguistic activities (**hugs*) and emoticons at the boundary of the sentence.

I would score Writer H as a Band 3 on the SWGDOC scale, which means it is highly probable that H is not X, as the text meets the criteria of substantial significant dissimilarities in range and variation, even though there were some similarities. If one had to assign a band based purely on the1,000-word level, then a Band 2 would be assigned due to the significant frequency difference regarding the *capital letters for emphasis* feature.

4.2.10 Summary of qualitative findings

In this section, all the candidate texts (Writers A-H) were analysed stylistically and individually against the disputed text X, with the purpose of finding the features shared with X and the features not shared with X at both the1,000-word and 2,000-word levels. Then SWGDOC band scales were assigned to help determine which candidate author was most probably also the author of the disputed text X. All the candidate authors shared features with X, as well as exhibiting various features that were not shared with X. Only Writer G exhibited meaningful similarities and very few dissimilarities, which led to the conclusion that it was highly probable that G is the writer of the disputed text X. A further observation was the difficulty in applying the SWGDOC band scales. The criteria descriptors lack the precision needed to reach an accurate decision. For example: *limitations are present* (Band 7) and *limitations may be present* (Band 6) leave a great deal of space for personal interpretation. Moreover, it appears that in this context *significant* should not be interpreted as a statistical term and so here too there is an element of subjectivity. This matter is discussed in more detail in 3.7.2.

A notable observation from the stylistic analysis is the effect of the text length. With the exceptions of Writers G and H, the text length played a minimal role as the1,000-word texts discriminated between the authors very similarly to the 2,000-word texts. Furthermore, at both levels substantial, evidence emerged in favour of Writer G being the author of the X text. The features I found to be most useful were the features not shared with Writer X, especially the idiosyncratic punctuation usage and DMC features.

The focus of the study will now shift to the study undertaken as a complementary investigation to the stylistic one, namely, the quantitative stylometric analysis.

4.3 Quantitative analysis

In this section I describe the findings from the quantitative analysis. I will begin with the keywords analysis, followed by function words, then the most frequently occurring words, and lastly punctuation. All the analyses are performed firstly at the1,000-word level and then at the 2,000-word level. The four sets of features are examined stylometrically with the focus being on the significant ($p \leq 0.05$) differences, which in relevant tables are shaded green, and very significant ($p \leq 0.01$) differences are shaded red. The *p*-values resulting from the Chi-square tests are displayed for each writer beneath the relevant column.

4.3.1 Keywords

This section examines the keywords at the1,000-word and 2,000-word levels between the reference corpus (X) and each of the study corpora in turn (A-H). The reader is referred to Chapter 3 section 3.9.1 for a detailed description of keywords. Referring to figure 4.2, which is a screenshot of the keyword analysis for Writer A, the column labelled *key word* shows the keywords for the text being examined and the second column shows the frequency counts for the keywords, with the third column displaying what percentage of the total words in the text a particular keyword represents. The column labelled *RC* refers to the reference corpus, which in this case is Writer X, and shows the frequency counts for the same keywords, but as they occur in the reference text. Many of the participants had numbers (ages, telephone numbers etc.) in their writings, and where numbers are key, they are represented by a single #.

Figure 4.2 - Writer A keywords

N	Key word	Freq.	%	RC.	RC. %	Keyness	P
1	I	45	4.63	16	1.58	14.46	0.00014
2	NOT	17	1.75	1	0.10	13.24	0.00027
3	AM	8	0.82	0		6.44	0.01116
4	WAS	10	1.03	2	0.20	4.40	0.03596
5	WILL	8	0.82	1	0.10	4.27	0.03886
6	HAD	8	0.82	1	0.10	4.27	0.03886
7	MY	13	1.34	4	0.40	4.13	0.04209
8	AT	3	0.31	16	1.58	-7.17	0.00739

The values in red at the bottom refer to keywords which have higher frequency counts in the reference corpus and are referred to as negatively key. The column labelled *keyness* gives a value to how key a word is. Kotzé (2010, 189) states that the higher the keyness value "the stronger the element of doubt as to common authorship". A further consideration is the ratio between grammatical and lexical keywords, as it has been argued that a high ratio of lexical keywords with low keyness values could imply common authorship (Kotzé 2010). The calculations for keyness and Chi-square are done automatically by WST. In the following sections, I discuss the keyword analyses for the1,000-word level and then the 2,000-word level. The screenshots of the keywords for Writers A – H are in Appendix 4. The findings for the keywords analysis at the1,000-word and 2,000-word level are presented using some of the criteria used by Kotzé (2010,190), as the average keyness values (the total of the values for all the keywords in a text, divided by the number of keywords), for example, are shown to "demonstrate the relatively high or low keyness value of items identified on the basis of their keyness" within a text.

4.3.1.1 Keywords (1,000-word level)

The results for the keyword analyses for Writers A-H are as follows:

Writer A/Writer X

Number of keywords:	8
Highest keyness value:	14.46
Lowest keyness value:	4.13
Average keyness value:	7.29
Ratio of grammatical words to lexical words:	8:0
Average keyness value of grammatical words:	7.29

The only keywords for Writer A were grammatical words, which are argued to be indicative of different grammatical vocabularies (Kotzé 2010). The number of keywords and the average keyness values are the highest in this group, so this would make A the least likely candidate for authorship of the X text.

Writer B/Writer X

Number of keywords:	6
Highest keyness value:	6.18
Lowest keyness value:	4.2
Average keyness value:	4.99
Ratio of grammatical words to lexical words:	3:3
Average keyness value of grammatical words:	4.79

The three lexical keywords (*Y, B* and *hey*) exhibited by Writer B were not used by Writer X. The *Y* is a DMC feature for *why* and *B* refers to the first letter of someone's name. The low number of grammatical keywords and their relatively low keyness values suggest that common authorship cannot be excluded.

Writer C/Writer X

Number of keywords:	6
Highest keyness value:	6.35
Lowest keyness value:	4.19
Average keyness value:	5.27
Ratio of grammatical words to lexical words:	4:2
Average keyness value of grammatical words:	5.27

The two lexical words (*ya, hey*) exhibited by Writer C were not used by Writer X. The relatively low number of keywords and keyness values means that common authorship cannot be excluded.

Writer D/Writer X

Number of keywords:	6
Highest keyness value:	7.13
Lowest keyness value:	4.10
Average keyness value:	5.3
Ratio of grammatical words to lexical words:	3:3
Average keyness value of grammatical words:	4.9

The two lexical words (*hey, company*) exhibited by Writer D were not used by Writer X. The '#' refers to numbers used by both writers. The relatively low number of keywords and keyness values means that common authorship cannot be excluded.

Writer E/Writer X

Number of keywords:	3
Highest keyness value:	5.17
Lowest keyness value:	4.18
Average keyness value:	4.54
Ratio of grammatical words to lexical words:	1:2
Average keyness value of grammatical words:	4.29

The two lexical words (*UK, hee*) exhibited by Writer E were not used by Writer X. The relatively low number of keywords and keyness values suggests that common authorship could be assigned.

Writer F/Writer X

Number of keywords:	7
Highest keyness value:	12.60
Lowest keyness value:	4.19
Average keyness value:	6.34
Ratio of grammatical words to lexical words:	4:3
Average keyness value of grammatical words:	6.9

Writer F exhibited three lexical words (*S, T* and *Dubrovnik*) and one grammatical word (*also*) not used by Writer X. S and T refer to the first letters of people's names and *Dubrovnik* is obviously a very specific geographic reference. The very high keyness value for the keyword *I* would suggest that there is not common authorship.

Writer G/Writer X

Number of keywords:	2
Highest keyness value:	4.12
Lowest keyness value:	4.02
Average keyness value:	4.07
Ratio of grammatical words to lexical words:	1:1
Average keyness value of grammatical words:	4.02

The only grammatical word that appears in both texts and that is also key for G is *we*, and this, given that it is a marginal keyword because it has a *p* value of 0.045, which is close to 0.05, would imply that Writers G and X use grammatical words similarly. The low keyness and very few keywords are a strong indication of common authorship.

Writer H/Writer X

Number of keywords:	6
Highest keyness value:	8.54
Lowest keyness value:	4.16
Average keyness value:	6.7
Ratio of grammatical words to lexical words:	4:2
Average keyness value of grammatical words:	7.2

The reasonably high keyness values exhibited for the keywords: *I, good, at* and *and* would suggest that there is no common authorship. More particularly, the average keyness value of grammatical words is the second highest in the group.

4.3.1.2 Keywords (2,000-word level)

The screen shots for keywords at the 2,000-word level are in Appendix 5 and the results of the keyword analyses are shown as follows:

Writer A/Writer X

Number of keywords:	14
Highest keyness value:	35.43
Lowest keyness value:	4.28
Average keyness value:	10.7
Ratio of grammatical words to lexical words:	11:3
Average keyness value of grammatical words:	12.5

The three lexical words (*God, Lord* and *church*) used by Writer A all refer to religion and are not exhibited in Writer X. For the grammatical keywords, the frequency counts are vastly different so as to suggest that common authorship is unlikely. For example, Writer A had 93 examples of *I* to Writer X's 29.

Writer B/Writer X

Number of keywords:	16
Highest keyness value:	12.77
Lowest keyness value:	4.35
Average keyness value:	6.99
Ratio of grammatical words to lexical words:	11:5
Average keyness value of grammatical words:	7.02

None of the lexical keywords exhibited by Writer B were used by Writer X. For the grammatical keywords, the frequency counts are considerably different, so as to suggest that common authorship is unlikely. For example, Writer B has 43 examples of *the* to Writer X's 70.

Writer C/Writer X

Number of keywords:	14
Highest keyness value:	13.54
Lowest keyness value:	3.88
Average keyness value:	6.6
Ratio of grammatical words to lexical words:	9:5
Average keyness value of grammatical words:	6.9

None of the lexical keywords exhibited by Writer C were used by Writer X. For the grammatical keywords, the frequency counts are considerably different and so to suggest that common authorship is unlikely. For example, Writer C has 84 examples of *you* to Writer X's 43.

Writer D/Writer X

Number of keywords:	21
Highest keyness value:	18.40
Lowest keyness value:	3.89
Average keyness value:	5.25
Ratio of grammatical words to lexical words:	12:9
Average keyness value of grammatical words:	6.3

Writer D had nine lexical keywords and of those nine, Writer X exhibited three, namely: numbers (#) (45/17), *years* (12/3) and *wow* (13/4). For the grammatical keywords, the frequency counts are considerably different, which would suggest that common authorship is unlikely. For example, Writer D has 82 examples of *you* to Writer X's 43.

Writer E/Writer X

Number of keywords:	15
Highest keyness value:	8.35
Lowest keyness value:	4.30
Average keyness value:	5.65
Ratio of grammatical words to lexical words:	10:5
Average keyness value of grammatical words:	5.6

None of the lexical keywords exhibited by Writer D were used by Writer X. For the grammatical keywords, the frequency counts are considerably different and suggest that common authorship is unlikely. For example, Writer E has 42 examples of *in* to Writer X's 25.

Writer F/Writer X

Number of keywords:	17
Highest keyness value:	24.7
Lowest keyness value:	3.97
Average keyness value:	8.00
Ratio of grammatical words to lexical words:	8:9
Average keyness value of grammatical words:	9.24

Writer F had eight lexical keywords relative to Writer X, and of that eight, Writer X exhibited only numbers (#) (17 of them, as opposed to F's four). For the grammatical keywords, the frequency counts are considerably different and suggest that common authorship is unlikely. For example, Writer F has 77 examples of *you* to Writer X's 43.

Writer G/Writer X

Number of keywords:	7
Highest keyness value:	7.78
Lowest keyness value:	4.05
Average keyness value:	5.25
Ratio of grammatical words to lexical words:	4:3
Average keyness value of grammatical words:	5.3

None of the lexical keywords exhibited by Writer G were used by Writer X. For the grammatical keywords, the frequency counts are not so different as to suggest that common authorship is unlikely. Writer G also has the lowest number of keywords and the lowest average keyness (shared with Writer D, though D has the highest number of keywords, at 21).

Writer H/Writer X

Number of keywords:	13
Highest keyness value:	24.77
Lowest keyness value:	4.15
Average keyness value:	7.54
Ratio of grammatical words to lexical words:	9:4
Average keyness value of grammatical words:	8.4

Writer H has four lexical keywords and the only shared lexical keyword was *good* (19/5). For the grammatical keywords, the frequency counts are considerably different and suggest that common authorship is unlikely. For example, Writer H has 81 examples of *I* to Writer X's 29.

4.3.1.3 Keywords conclusion

Table 4.10 gives an overview of the number of keywords and average keyness for the1,000-word and 2,000-word levels.

Table 4.10 – Keyword summary (1,000-word and 2,000-word levels)

Suspect/Author	Keywords: 1,000 words	Average keyness: 1,000 words	Keywords: 2,000 words	Average keyness: 2,000 words
Writer X/Writer A	8	7.29	14	10.7
Writer X/Writer B	6	4.99	14	6.99
Writer X/Writer C	6	5.27	16	6.6
Writer X/Writer D	6	5.3	21	5.25
Writer X/Writer E	3	4.54	15	5.65
Writer X/Writer F	7	6.34	17	8
Writer X/Writer G	**2**	**4.07**	**7**	**5.25**
Writer X/Writer H	6	6.7	13	7.54

The above table clearly shows that Writer G is correctly clustered with Writer X, having the lowest keywords and average keyness at both the1,000-word and 2,000-word levels. However, there are a number of red flags, particularly at the1,000-word level. Writer E, for example, has three keywords and an average keyness of 4.54, which is close to G. Moreover, although G also has the lowest average keyness of grammatical words (4.07), it is E that has the second lowest (4.54). Thus some limitations to the discriminating power of these two style markers are revealed at this level.

The results at the 2,000-word level are less ambiguous. Despite G having to share the lowest keyness value of 5.25 with D, G has the lowest number of keywords (7), just above half the number exhibited by H, its nearest other possible candidate author in terms of this style marker. Also, G again has the lowest in terms of average keyness of grammatical words (5.3), and although it is again E that has the second lowest (5.6), E's candidature as Writer X is greatly weakened by the fact that her text shows up no fewer than 15 keywords. It seems that the three most important features in the keyword analyses are quite effective at signalling common authorship and that, for my study, the number of keywords is the most effective of the three style markers.

The keyword test highlights the importance of text size. A lower text size results in lower frequency counts for individual words, which affects the statistical measurements, as there will not be sufficient numbers to calculate statistical significance. For that reason, I would adduce that the results from the keyword test for a text size of 1,000 words or less should be treated with a fair degree of caution.

Kotzé (2010, 192) suggests that for his study an average keyness value of 10 for (preferably) grammatical words would be a "reliable cut-off point regarding the keyness of a comparison between two documents", implying that only above this level could one be reasonably sure that the document had different authors. Kotzé (2010) noted in his investigation of the *Father Punch* case that in comparing texts written by the suspect, the keyness value never rose above 15 and if those texts had been compared to texts by another author, they would have exceeded 15, thus leading to the conclusion that the higher the keyness value the lower the likelihood of common authorship. However, due to a variety of textual features this cut-off point can be re-evaluated. Kotzé (2010, 192) states that: "The measure of confidence with which a conclusion could be made depends on a number of variables including, for example, the size of the text and textual density", and for these reasons the cut-off point can vary.

This is shown by the fact that the cut-off point of 10 would not be realistic for my study as all my participants had a lower than 10 average at the1,000-word level and all but Writer A at the 2,000-word level. It should be noted that the *Father Punch* core document consisted of 5,482 words and the 11 chronicles amounted to 25,431 words. Kotzé (2010) does not give a word count for the *Angry Academic* case, which resulted in the cut-off point of 10. My deduction is that the length of text greatly influences the cut-off point and the determination of the cut-off point should be on a case by case basis.

4.3.2 Function words

This section presents the findings of the analysis of the function words as postulated by Morton (1978) at both the 1,000-word and 2,000-word levels.

4.3.2.1 Function words (1,000-word level)

The results of the analysis comparing the frequencies of function words is set out in table 4.11 at the1,000-word level. In table 4.11, the disputed text (X) is compared to the eight other authors of similar demographic backgrounds. A look at the table shows that G and E have no shaded blocks, meaning that there are no significant differences in the use of those function words between G and X and E and X. Writers B and D only have one shaded block apiece, C and F have two, and all the remaining authors have three shaded blocks.

Table 4.11 – Function words (1,000-word level)

		X	A	B	C	D	E	F	G	H
1	**A**	**21**	13	29	24	29	16	14	23	10
2	**All**	**2**	7	2	7	0	5	8	5	8
3	**Also**	**0**	3	1	3	1	3	6	5	2
4	**And**	**36**	27	35	16	25	34	23	24	12
5	**Any**	**0**	3	0	2	0	1	1	0	1
6	**As**	**6**	4	5	1	4	6	9	5	1
7	**At**	**16**	3	7	4	7	7	4	12	3
8	**Been**	**7**	1	12	3	14	3	3	6	3
9	**But**	**5**	8	10	7	6	6	8	6	3
10	**For**	**10**	12	11	14	12	7	6	15	9
11	**In**	**15**	15	10	11	26	26	12	13	8
12	**It**	**9**	3	10	9	9	8	17	12	14
13	**No**	**5**	5	1	4	1	3	4	2	5
14	**Not**	**1**	16	2	8	6	3	5	2	3
15	**Of**	**13**	11	8	10	11	6	15	5	9
16	**On**	**2**	9	12	8	10	7	7	5	9
17	**That**	**7**	14	4	10	6	7	8	4	5
18	**The**	**26**	43	33	16	24	29	26	31	27
19	**This**	**6**	1	7	4	1	3	3	5	9
20	**To**	**29**	37	27	26	26	22	45	26	38
21	**Very**	**4**	3	1	4	2	3	4	1	2
22	**Was**	**2**	10	5	8	5	4	11	5	4
23	**Were**	**1**	2	2	0	0	0	0	5	3
	p		0.00	0.17	0.02	0.14	0.45	0.00	0.76	0.00

Turning now to the significance levels between the profiles of A to H compared to X, it can be seen that writers B, D, E and G meet the requirements of the null hypothesis of sameness as they all exhibit $p>0.05$ (Chaski 2001). However, a closer analysis reveals some considerable differences. In order of furthest to closest to X, it can be seen that D is the furthest away from X, followed by B and E. G is the closest to X with a *p* value of 0.76.

4.3.2.2 Function words (2,000-word level)

In table 4.12, the disputed text (X) is compared with the others at the 2,000-word level. It can be seen that each author has some shaded blocks, even those which did not at the1,000-word level.

Table 4.12 – Function words (2,000-word level)

		X	A	B	C	D	E	F	G	H
1	**A**	**36**	37	55	41	50	42	36	39	27
2	**All**	**6**	8	6	13	4	8	12	9	15
3	**Also**	**1**	3	3	3	2	4	9	12	2
4	**And**	**61**	58	61	33	43	54	56	56	36
5	**Any**	**6**	3	1	3	1	3	3	2	2
6	**As**	**11**	15	8	8	6	10	13	7	3
7	**At**	**28**	4	12	9	15	11	5	19	7
8	**Been**	**12**	4	17	8	17	5	7	8	4
9	**But**	**11**	14	15	11	11	11	9	15	13
10	**For**	**26**	23	23	31	20	23	18	21	20
11	**In**	**25**	33	19	25	46	40	31	27	28
12	**It**	**21**	13	15	25	11	22	32	24	27
13	**No**	**5**	9	5	4	2	4	4	2	6
14	**Not**	**5**	29	4	12	9	6	9	5	7
15	**Of**	**31**	32	20	22	20	20	30	21	18
16	**On**	**8**	18	20	18	22	12	18	10	14
17	**That**	**12**	25	18	19	12	21	18	8	15
18	**The**	**71**	71	64	43	50	60	44	60	55
19	**This**	**11**	8	13	7	3	7	12	9	13
20	**To**	**63**	72	57	44	55	40	85	74	67
21	**Very**	**8**	4	2	5	2	6	4	1	6
22	**Was**	**3**	15	17	16	6	12	15	10	6
23	**Were**	**3**	2	6	2	0	1	0	5	3
	p		0.00	0.01	0.00	0.00	0.00	0.00	0.16	0.00

Unlike the1,000-word level, where four of the eight writers met the $p>0.05$ criteria, only G met this criterion at the 2,000 word level with a p value of 0.16. This style marker was quite effective at identifying the author of the ‘disputed text’. Writer G had the most similar profile to X at both the1,000-word and 2,000-word levels, and that was despite the strong drop from 0.76 at the1,000-word level to 0.16 at the 2,000-word level. This

drop was due partly to G's consistent high use of *also* and *was* into the second 1,000 words, while in the X text they remained at very low levels. On the other hand, the X text continued to show *very* quite often in the second half, while there was no addition by G of this word. The reasons why such differences arose are beyond the scope of this study.

4.3.3 Most frequently occurring words

This section will examine the findings for the 30 most frequently occurring words from X's profile, and how they appear in the other eight authors' writings at both the1,000 and 2,000-word level.

4.3.3.1 Most frequently occurring words (1,000-word level)

The results of the analysis comparing the frequencies of the most commonly occurring lexical items in X's submission at the1,000 word level are set out in Table 4.13. A look at the table shows that G has no shaded blocks, which means that there are no significant differences in the use of those most frequently occurring words between G and X. The blocks shaded blue indicate the lexical items which are shared with the function word test.

Considering the significance levels between the profiles of Writers A-H as compared to X, it can be seen that apart from G, E's *p* value of 0.48 also meets the requirements of the null hypothesis of sameness as they both exhibit $p>0.05$ (Chaski 2001) and C nearly does (at 0.05 exactly). On this basis, then, E should not be excluded as X, but with a *p* value of 0.86, G's profile is the closest to X. It can be seen that writers with the highest *p* values have the least number of significant and very significant figures.

Table 4.13 – Most frequently occurring (1,000 word level)

		X	A	B	C	D	E	F	G	H
1	**A**	**21**	13	29	24	31	16	14	25	11
2	**About**	**7**	5	0	9	6	4	4	3	4
3	**And**	**36**	27	35	17	25	34	23	26	14
4	**Are**	**10**	4	8	6	20	12	7	10	8
5	**As**	**7**	4	5	2	4	6	10	5	1
6	**At**	**16**	3	7	4	7	7	4	25	3
7	**Been**	**7**	1	13	3	14	3	3	6	3
8	**Do**	**7**	5	3	6	1	7	10	5	5
9	**Going**	**7**	2	2	4	1	2	3	6	7
10	**For**	**11**	12	11	13	12	8	6	16	10
11	**From**	**6**	3	4	2	2	0	3	2	5
12	**Have**	**13**	8	9	10	6	9	12	11	8
13	**Here**	**8**	0	1	2	2	7	3	1	2
14	**How**	**6**	3	3	6	14	4	7	6	0
15	**I**	**16**	45	18	29	25	27	44	12	38
16	**I'm**	**6**	0	7	1	5	7	6	4	16
17	**In**	**15**	15	10	11	27	26	12	13	13
18	**Is**	**15**	14	8	9	5	13	12	7	12
19	**It**	**9**	4	10	8	9	9	17	13	14
20	**It's**	**7**	0	3	0	5	4	0	6	5
21	**Like**	**12**	1	3	2	1	3	0	6	5
22	**Love**	**6**	4	2	2	0	3	4	5	0
23	**Of**	**13**	11	8	10	11	7	15	5	9
24	**Really**	**7**	3	3	0	1	0	4	3	4
25	**That**	**7**	14	4	11	6	7	8	4	5
26	**The**	**28**	43	33	16	24	30	27	31	29
27	**To**	**29**	38	28	26	26	23	46	26	34
28	**With**	**11**	3	0	11	8	12	6	8	7
29	**You**	**23**	19	40	29	45	33	27	26	35
30	**Your**	**12**	7	6	8	2	10	7	10	7
	p		0.00	0.02	0.05	0.00	0.48	0.00	0.86	0.00

4.3.3.2 Most frequently occurring words (2,000-word level)

The results of the analysis comparing the frequencies of the most commonly occurring items in Writer X's submission at the 2,000 word level are set out in table 4.14. A look at the table shows that Writer G again has no shaded blocks, which means that there are again no significant differences in the use of those most frequently occurring words between Writer G and Writer X.

In dramatic contrast to the 1,000 word table, it can be seen that only writer G meets the criterion of $p>0.05$ with what Olsson (2004) and Grant and Baker (2001) would regard as an 88% probability of being the author of the disputed text. G is the only writer to have no significant or very significant differences and met the $p>0.05$ criteria.

Table 4.14 - Most frequently occurring words (2000 word level)

		X	A	B	C	D	E	F	G	H
1	A	36	39	55	41	50	42	36	39	28
2	About	19	10	3	15	10	6	8	12	7
3	And	61	58	61	34	43	54	57	56	36
4	Are	25	10	20	8	30	19	13	16	11
5	As	11	15	10	9	6	10	13	7	3
6	At	28	4	12	9	15	11	5	19	6
7	Be	14	16	12	9	15	9	20	11	13
8	Been	12	4	17	9	17	5	8	8	4
9	For	26	23	23	30	20	23	19	21	20
10	Going	12	2	6	8	2	5	3	7	15
11	Have	22	18	25	19	14	22	21	23	14
12	I	29	93	33	61	41	53	82	32	81
13	I'm	12	0	15	6	16	11	7	10	27
14	In	25	33	19	25	46	42	32	27	28
15	Is	21	35	19	22	16	20	23	19	24
16	It	21	14	15	24	11	22	32	24	27
17	It's	16	0	7	3	9	7	0	19	16
18	Like	20	4	7	3	4	6	0	11	11
19	Love	14	10	5	2	2	12	11	11	2
20	Me	12	19	11	14	23	14	13	10	17
21	Of	31	32	20	22	20	21	31	21	20
22	Really	15	5	4	2	2	2	9	8	6
23	So	17	8	28	24	17	10	7	20	21
24	That	12	25	19	19	13	21	18	8	16
25	The	70	71	64	43	50	61	45	60	55
26	To	64	73	58	45	55	40	85	75	69
27	What	16	8	14	2	16	12	7	11	8
28	With	15	12	20	20	16	23	19	14	12
29	You	43	42	84	52	82	69	77	56	62
30	Your	23	15	19	11	9	14	11	19	18
	P		0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.00

4.3.4 Punctuation

This section presents the findings of analysis of the punctuation used by the writers at both the1,000-word and 2,000-word levels.

4.3.4.1 Punctuation (1,000-word level)

The results of the analysis comparing the frequencies of the different punctuation items at the1,000-word level are set out in Table 4.15.

The most noticeable aspect of this table is the different types of punctuation employed by the eight authors, from the standard prescriptive use of punctuation to the more creative examples, such as using multiple question marks and emoticons. Many of the features were only used by one or two authors, which resulted in many frequency counts being too low to be statistically viable. Despite the abundance of non-standard examples, all the authors employed features of standard punctuation. It can be seen that G has the fewest shaded blocks, with just one shaded feature (‘dash’) used in a significantly different manner to X.

Turning now to the significance levels between the profiles of writers A-H compared to X, it can be seen that only G met the $p>0.05$ criterion (Chaski 2005) with a p value of 0.46. The profiles of the others were very significantly different to X.

Table 4.15 – punctuation (1,000-word level)

Feature	e.g.	X	A	B	C	D	E	F	G	H
Full stop	.	**61**	29	30	38	17	26	61	58	33
Comma	,	**15**	9	23	17	27	9	43	15	49
Apostrophe (possession)	‘	**3**	1	3	0	0	0	0	3	1
Apostrophe (omissions)	‘	**45**	5	30	15	27	41	40	52	57
Colon	:	**0**	2	1	1	1	0	0	1	0
Question mark	?	**19**	3	25	18	33	33	4	22	12
Exclamation mark	!	**11**	3	10	31	30	4	1	10	15
Dash	–	**10**	46	3	37	0	20	10	3	0
Hyphen	-	**1**	2	0	1	0	1	0	1	0
Slash	/	**1**	1	1	0	3	1	0	1	1
Single quotes*	‘...’	**4**	4	1	0	1	0	1	0	1
Double quotes*	“...”	**0**	3	0	0	0	0	0	0	1
Brackets*	(...)	**2**	10	3	2	0	12	1	4	0
Ellipsis (2 dots)*	..	**0**	0	0	2	32	0	5	0	0
Ellipsis (3 dots)*	...	**0**	1	20	1	45	27	0	2	13
Ellipsis (4 dots)*	....	**0**	0	2	0	2	15	1	9	5
Ampersand	&	**0**	0	0	0	4	0	0	0	0
Asterisks*	*.....*	**0**	0	0	1	0	0	0	0	1
At symbol	@	**1**	0	0	0	0	0	1	0	0
Multiple Q. Marks – 2*	??	**0**	1	0	1	18	0	0	0	1
Multiple Q. Marks – 3*	???	**0**	0	3	0	2	1	0	0	0
Multiple Q. Marks – 4*	????	**0**	2	2	0	0	0	0	0	0
Multiple E. Marks – 2*	!!	**0**	0	0	4	15	2	5	0	4
Multiple E. Marks – 3*	!!!	**0**	0	7	0	17	6	1	0	0
Multiple E. Marks – 4*	!!!!	**0**	0	0	0	2	1	1	0	0
Multiple E. Marks*	!!!!!!!	**0**	0	0	0	1	1	0	0	0
Q. Mark + E. Mark*	?!	**0**	0	0	0	0	0	0	2	0
E. Mark + Q. Mark	!?	**0**	0	0	0	17	0	0	0	1
Emoticons	☺	**0**	0	0	2	0	0	0	0	0
Emoticons	:-)	**0**	0	0	0	0	0	3	0	0
Emoticons	:)	**0**	0	2	18	8	0	13	0	4
Emoticons	:(	**0**	0	1	0	3	0	0	0	1
Emoticons	:D	**0**	0	0	0	1	0	1	0	4
Emoticons	:P	**0**	0	0	0	1	0	0	0	1
Emoticons	;)	**0**	0	0	0	1	0	0	0	1
	p		0.00	0.01	0.00	0.00	0.00	0.00	0.46	0.00

* Denotes pairs or multiples of features that are counted as singular.

4.3.4.2 Punctuation (2,000-word level)

The results of the analysis comparing the frequencies of the different punctuation items at the 2,000 word level are set out in Table 4.16.

Just as in the1,000-word level, there is considerable variety in the punctuation features employed by the eight writers. It can be seen that G has no red or green shaded blocks, meaning that she has not used those punctuation marks in any significantly different way to X. An examination of the significance levels between the profiles of writers A-H compared to X showed that only G met the $p>0.05$ criterion with a p value of 0.60.

Table 4.16 – punctuation (2,000-word level)

Feature	e.g.	X	A	B	C	D	E	F	G	H
Full stop	.	**107**	56	93	86	28	30	123	109	71
Comma	,	**26**	29	67	32	50	11	95	30	91
Apostrophe (possession)	‘	**3**	2	7	0	4	1	0	3	1
Apostrophe (omissions)	‘	**94**	12	78	36	68	71	84	110	113
Colon	:	**2**	2	1	1	0	0	1	1	0
Question mark	?	**47**	7	53	29	29	76	10	45	24
Exclamation mark	!	**16**	7	19	51	46	10	2	15	25
Dash	–	**9**	54	7	54	0	48	13	6	0
Hyphen	-	**10**	2	1	2	2	4	0	6	3
Slash	/	**2**	1	1	0	4	1	2	3	0
Single quotes*	‘...’	**7**	3	1	0	6	2	1	2	2
Double quotes*	“...”	**0**	3	0	1	0	0	0	0	1
Brackets*	(...)	**3**	16	4	4	4	18	2	7	8
Ellipsis (2 dots)*	..	**0**	0	0	10	8	1	0	0	0
Ellipsis (3 dots)*	...	**2**	5	50	6	110	44	0	4	29
Ellipsis (4 dots)*	....	**5**	3	4	2	0	25	0	11	7
Ampersand	&	**0**	0	0	0	1	0	0	0	2
Asterisks*	*.....*	**0**	0	0	1	0	0	0	0	2
At symbol	@	**2**	0	0	0	0	0	1	0	0
Multiple Q. Marks – 2*	??	**0**	1	0	3	26	1	1	0	8
Multiple Q. Marks – 3*	???	**0**	0	8	0	5	4	0	0	0
Multiple Q. Marks – 4*	????	**0**	2	3	0	0	2	0	0	0
Multiple E. Marks – 2*	!!	**2**	0	0	6	36	2	5	0	0
Multiple E. Marks – 3*	!!!	**0**	4	14	2	34	8	1	0	0
Multiple E. Marks – 4*	!!!!	**0**	4	1	0	4	4	1	0	0
Multiple E. Marks*	!!!!!!	**0**	3	0	0	5	1	0	0	0
Q. Mark + E. Mark*	?!	**0**	0	0	0	0	0	0	2	4
E. Mark + Q. Mark	!?	**0**	0	0	0	16	0	0	0	1
Emoticons	☺	**0**	0	0	2	0	0	13	0	0
Emoticons	:-)	**0**	0	0	0	0	0	2	0	0
Emoticons	:)	**0**	0	4	31	16	0	12	0	4
Emoticons	:(	**0**	0	0	2	3	0	0	0	2
Emoticons	:D	**0**	0	0	0	1	0	1	0	6
Emoticons	:P	**0**	0	0	0	2	0	0	0	1
Emoticons	;)	**0**	0	1	0	8	0	0	0	2
	p		0.00	0.00	0.00	0.00	0.00	0.00	0.60	0.00

***** Denotes pairs or multiples of features that are counted as singular.

4.4 Summary of results

This section sets out a summary of the qualitative and quantitative findings at firstly, the1,000-word level and secondly, the 2,000-word level.

4.4.1 Qualitative summary (1,000-word level)

The qualitative analysis involved comparing stylistic features in texts written by Writers A-H to the text written by Writer X. Despite the relatively small text size of1,000 words, it was possible to discern discriminating features between the texts and to reach a conclusion using the SWGDOC band scales. Table 4.17 gives an overview of the findings.

Table 4.17 – Stylistic findings (1,000-word level)

Writer	SWGDOC Score	Features shared / unshared	Significant differences in shared features	Significant differences in unshared features
A	2	5 / 4	Word+space+dash+space +word	Non-standard word initial capitalisation
B	2	3 / 7	-	Word+no space+ellipsis dots
C	1	2 / 9	Word+space+dash+space +word	Lowercase *i* for 1st person singular Emoticons
D	2	3 / 10	-	Word+no space+ellipsis dots Non standard punctuation !? ?? !! !!!
E	2	4 / 6	Word+space+dash+space +Word	Word+no space+ellipsis+ space
F	3	4 / 4	-	-
G	7	5 / 2	-	-
H	2	4 / 9	-	-

4.4.2 Qualitative summary (2,000-word level)

The qualitative analysis at the 2,000-word level showed promising results, and it was possible to discern discriminatory features between the texts and to reach a conclusion using the SWGDOC band scales. Table 4.18 gives an overview of the findings.

Table 4.18 – Stylistic findings (2,000-word level)

Writer	SWGDOC Score	Features shared / unshared	Significant differences in shared features	Significant differences in unshared features
A	2	5 / 4	Capital letters for emphasis Word+space+dash+space +word Brackets	Non-standard word initial capitalisation
B	2	3 / 7	-	Word+no space+ellipsis dots
C	1	3 / 9	Word+space+dash+space +word	Apostrophe omissions Lowercase *i* for 1st person singular Emoticons
D	2	3 / 11	-	Word+no space+ellipsis dots Non standard punctuation !? ?? !! !!!
E	2	4 / 6	Word+space+dash+space +Word Brackets	Word+no space+ellipsis+ space Word+no space+ellipsis dots
F	3	4 / 4	Word+space+dash+space +word	Sentence ending emoticons
G	8	7 / 2	Capital letters for emphasis	-
H	3	4 / 9	-	Word+no space+ellipsis+ Space

The analysis of the stylistic features identified G as the writer of the ‘disputed’ text and the most effective style marker was the idiosyncrasies associated with punctuation. A further observation was that text size had a minimal effect, with the1,000-word level being as good as the 2,000-word level.

4.4.3 Quantitative summary (1,000-word level)

Table 4.19 shows the aggregate results of the three tests (function words, most frequently occurring words and punctuation) which were subjected to the Chi-square test at the 1,000-word level. The blocks shaded light purple indicate the writers which meet the criterion of $p>0.05$.

Table 4.19 – Aggregate results (1,000-word level)

	A	B	C	D	E	F	G	H
Function words	0.00	0.17	0.02	0.14	0.45	0.00	0.76	0.00
Frequently occurring words	0.00	0.02	0.05	0.00	0.48	0.00	0.86	0.00
Punctuation	0.00	0.01	0.00	0.00	0.00	0.00	0.46	0.00
Aggregate	**0.00**	**0.06**	**0.02**	**0.04**	**0.31**	**0.00**	**0.69**	**0.00**

The function word style marker was the least successful in identifying the author of the disputed text with four of the writers (B, D, E and G) exceeding the criterion of $p>0.05$. The frequently occurring words style marker was more successful with only two writers (E and G) reaching the $p>0.05$ criterion. In both cases, Writer G's p values far exceed those of E, which would imply that the two tests have correctly identified the author of the disputed text, but with some element of doubt as E's p values are quite high. Punctuation, on the other hand, has correctly identified G as the author of the disputed text by a reasonable margin. My concern is whether G's p values are acceptable for a court of law. Both Olsson (2004), and Grant and Baker (2001) have raised doubts about $p>0.05$ being sufficient to infer that two texts have the same author, and Grant and Baker (2001, 76) have stated that "perhaps the only way we could accept that two samples are really from the same population is when the probability level is more than 0.95". G's p values of 0.76, 0.86 and 0.46 do not come close to Grant and Baker's (2001) threshold of $p\geq0.95$.

4.4.4 Quantitative summary (2,000-word level)

Table 4.20 shows the aggregate results of the three tests which were subjected to the Chi-square statistic at the 2,000 word level. Again, the blocks shaded light purple indicate where the criterion of $p>0.05$ is met.

Table 4.20 – Aggregate results (2,000-word level)

	A	B	C	D	E	F	G	H
Function words	0.00	0.01	0.00	0.00	0.00	0.00	0.16	0.00
Frequently occurring words	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.00
Punctuation	0.00	0.00	0.00	0.00	0.00	0.00	0.60	0.00
Aggregate	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.54**	**0.00**

The 2,000 word table shows a remarkably different result to the1,000-word table. Only G meets the criterion of $p>0.05$ for all three style markers. Despite correctly identifying G as the author of the disputed text, function words are the least successful of the three. At the 2,000-word level, the most frequently occurring words have proven to be the most accurate, followed closely by punctuation. Just like the 1,000-word level, there is the concern that G's *p* values do not reach Grant and Baker's (2001) threshold of 0.95.

4.5 Conclusion

My study has two aims: (1) The first specific aim is to test the extent to which each member of a set of select style markers can effectively identify the writer of a disputed text, within the (sub) genre of Facebook using both stylistic and stylometric approaches. (2) The second specific aim is to determine if the length of text i.e.1,000-word and 2,000 -word levels, has any bearing on the efficacy of the chosen style markers.

The qualitative analysis revealed a number of differences between the1,000-word and 2,000-word levels. The selection of the style markers to be analysed followed a bottom up approach. The disputed text X was analysed for stylistic features which could be considered marked and the following seven features were chosen: (1) *Capital letters for emphasis,* (2) *words/phrases highlighted in single quotes,* (3) *word+space+dash+ space+word,* (4) *word+no space+ellipsis (4 dots)+no space+word,* (5) *two or more*

questions in a row, (6) *brackets to show afterthoughts or to give extra information,* and (7) *apostrophe+cause*. Thereafter, each of the writers' texts were analysed for those seven style markers. In addition to those seven stylistic features, each text was analysed for features not found in the disputed text X and were categorised according to the following criteria: (1) *punctuation, typography and spelling,* (2) *lexis,* and (3) *digitally mediated communication features*. In order to give my conclusions more scientific validity, I used a set of band scales developed by the Scientific Working Group for Forensic Documents Analysis (SWGDOC). The SWGDOC scales have specific criteria which allow one to award a text a band score between 1 and 9. A Band 1 would imply that the texts are so different as to eliminate any possibility of common authorship and Band 9 means a definite common authorship.

At the 1,000-word level, all the writers shared some features with Writer X, though sometimes in significantly different numbers, and all the writers except Writer G exhibited substantial numbers of features not shared with Writer X. Therefore, based on the number of similarities and absence of dissimilarities, Writer G was the only writer who met the criteria for scoring 'highly probable' on the SWGDOC scales.

At the 2,000-word level, all the writers shared some of the features with the disputed text X, as well as having unshared features. However, only Writer G shared all seven features with Writer X, and the only unshared features were a few non-standard typographical renditions and lexical choices. Therefore, based on the number of similarities and minor dissimilarities, Writer G was again the only writer who met the criteria for scoring 'highly probable' on the SWGDOC scales. A full Band 9 was not awarded due the significant differences in the use of *capital letters for emphasis,* and minor typographical renditions and non-standard lexis.

The quantitative analysis revealed marked differences between the1 000-word and the 2,000-word levels. The style markers selected for the stylometric analyses were: (1) *keywords,* (2) *Morton's (1978) function words*, (3) *most frequently occurring words* and

(4) *punctuation.* It is in the quantitative section where the greatest differences in results between the1,000-word level and 2,000-word level can be found.

At the 1,000-word level, the keyword test accurately clustered Writer G with Writer X (two keywords and an average keyness of 4.07). However, this result is not without a shadow of doubt as Writer E's keyword analysis result was not that dissimilar (three keywords and an average keyness of 5.54). At the 2,000-word level, the keyword test again clustered Writer G as the most likely author of the disputed text with a more robust display of results. Writer G had seven keywords with an average keyness of 5.25. The keyword analyses followed the methodology used by Kotzé (2010). The three main keyness measures successfully indicated that G was the author of the disputed text, but at the 2,000-word level it was the number of keywords that appeared to be the most effective.

Morton's (1978) list of function words were originally selected for stylometric analyses conducted on literature as Morton claimed they were the most commonly used function words. They were used in a forensic linguistic context by Hubbard in 1995 in dealing with extortion letters made to a large supermarket chain in South Africa. In my study, at the 1,000 word level, these function words were a little problematic, because where the frequencies of a feature for a known writer and Writer X together did not reach a combined count of ten, that feature was eliminated from the chi-square calculations for that writer. The results at the1,000-word level were partially conclusive as four of the writers had *p* values equal to or above 0.05. However, Writer G had the highest *p* value (0.76), which was quite considerably higher than second place Writer E (0.45). This was the least successful discriminating test, and that could be due to the top-down nature of imposing features which were not likely to have high frequency counts because of the differences in genre between literature and social networking. At the 2,000-word level, only G reached the $p>0.05$ threshold with a *p* value of 0.16, a substantial drop from the 1,000-word level result, which was attributed to the higher frequency counts of *also* and *was* in the second set of 1,000 words.

When compared with Morton's function words, the analysis of the most frequently occurring words (29 of the 30 words were function words) yielded better results at both the 1,000-word and 2,000-word levels. At the1,000-word level, the negative effects of a low word count are still evident, with two of the eight writers scoring above 0.05. Writer G had the highest *p* value of 0.86, followed by writer E with 0.48. At the 2,000-word level, the picture is remarkably different in that only Writer G meets the requirement of $p>0.05$, with a result of 0.88. This test was more successful than the function words test as the automatic text-based approach to selecting the words meant that all the words were likely to be exhibited in the majority of texts.

Punctuation formed part of both the qualitative and quantitative analyses. In the stylistic analysis, punctuation was judged mostly in terms of idiosyncratic uses that were noticed. The nature of digitally mediated communication for social purposes is very tolerant of what would be considered non-standard punctuation practices. Different punctuation habits helped discriminate between writers. In the quantitative analysis, the focus shifted to counting punctuation features, unusual or not, similarly to work conducted by Chaski (2001 and 2005). Looking at the frequency counts, one can see that the standard punctuation usage dominates, whilst the creative punctuation features are typical to specific writers, where their discriminatory power in the qualitative assessment becomes apparent. At the1,000-word level, the punctuation test was the most successful with Writer G being the only author to exceed the $p>0.05$ threshold with a *p* value of 0.46. The punctuation test at the 2,000-word level showed the most discriminating power, again with only Writer G exceeding the threshold with a *p* value of 0.60. Of the three tests, punctuation proved to be the style marker with the most discriminating power and fulfils the requirements of this study's first aim of exploring the extent to which each member of a set of style markers can effectively identify the writer of a disputed text.

The above paragraphs return us to the question of what effect the length of the text has on the validity of the chosen style markers. Morton (1978, 209) has stated that "the omens are favourable" if one has 2,000 or more words, and "in many cases a sample of

one thousand words will be adequate, but with less than that complications are to be expected" (1978, 200). It can be seen that 1,000 words may not be completely adequate if one is doing a stylometric analysis and the results should be treated with some caution, particularly in light of the fact that other writers also reached the $p>0.05$ threshold. This is probably due to the fact that frequency counts require a sizeable number of examples in order to reach conclusions of significant differences, and 1,000 words may not be sufficient for a conclusive result. Chaski (2011), in an Internet discussion forum regarding the Zuckerberg/Ceglia case investigated by McMenamin (see Liberman 2011), recommends 2,000 words or 100 sentences per author to obtain the most optimal results. The results at the 2,000-word level for all four tests were more conclusive. However, arguably the stylistic analyses proffered more robust results at the1,000-word level and those results were reinforced at the 2,000-word level. With reference to Kotzé's (2010) quote at the beginning of this chapter, we need to be absolutely sure of our conclusions, and stylometric results at the1,000-word level may not always offer that peace of mind.

Chapter 5 – Conclusion

> *In unselfconscious utterance, certain features occur—relatively permanent features of the speech or writing habits—which identify someone as a specific person, distinguishing him from other users of the same language.*
>
> (Crystal and Davy 1969, 66)

5.1 Introduction

In this research study I have endeavoured to extend the use of stylistic and stylometric tools to the realm of Facebook and social networking in order to test known forensic linguistic methods of authorship attribution in the newest and fastest growing medium of communication. This chapter begins by reviewing the previous four chapters before moving on to briefly assess the contribution made by this study. The chapter then moves on to the limitations of the study and finally there are suggestions for future research.

5.2 Overview of the study

This section presents an overview of the preceding four chapters, starting with a restatement of the aims of the study and how they were to be achieved.

The overall aim of this study was to explore to what extent it is possible to attribute authorship on a publicly accessible social networking site such as Facebook to a single author among a group of authors from similar demographic backgrounds. This aim was to be achieved in terms of two specific aims. Aim 1 was:

To explore the extent to which each member of a set of style markers can effectively identify the writer of a disputed text.

This aim was in turn to be achieved by way of two objectives, namely: (1) analysing the texts stylistically by comparing in particular the differences and similarities in spelling, punctuation and typography; lexis; and DMC features (aspects of layout were also looked at, as was grammar, but for reasons explained in Chapter 3, they were not given

systematic consideration in this study); and (2) subjecting the texts to stylometric analyses of four style markers, namely: keywords; function words; most frequently occurring words; and punctuation.

The second aim (Aim 2) was:

To explore the extent to which the length of texts available is a factor in how effectively the writer of a disputed text can be identified.

Two thousand words of text is recommended for stylometric analyses (Chaski 2011), but Morton (1978) stated that 1,000 words would be adequate, although anything less would result in complications. Therefore, stylistic and stylometric analyses were done at the 1,000-word and 2,000-word levels, followed by a comparison of the results to determine the effects of text size.

Chapter 1 began by examining the history and intricate workings of Facebook and social networking sites. Then, the language used on Facebook and social networking sites was discussed, using examples from my participant authors in conjunction with theoretical research conducted by Danet (2001) and Crystal (2007 and 2011). Thereafter, Facebook and social networking sites were situated into their appropriate sociocultural context for a forensic linguistic study and noting that a forensic linguistic research gap exists in this context by examining criminal activity perpetrated using Facebook and social networking sites (boyd and Ellison 2007 and Olsson 2010), so providing the *raison d'être* for the study.

Chapter 2 looked at the literature and research relevant to my study. The chapter began by examining the history of forensic linguists and stylometrics from the early disputes over the authorship of the *Iliad* and the *Odyssey,* and the works of Shakespeare before moving to more recent examples. The review examined a number of pertinent recent cases, namely: The Evans Statements (Coulthard and Johnson 2007), The Derek Bentley Case and The Unabomber (Coulthard 2000), as well as a notable problematic

attempt of stylometry, namely The CUSUM Method (Juola 2006), which was highlighted to show how this field of study is evolving. Thereafter, I examined concepts relevant to forensic linguistics, with a strong emphasis on the idiolect debate, which drew on research conducted by McMenamin (2002), Coulthard (2004), Olsson (2008) and Grant (2010). In order to fully grasp the reasons why and how language used in digitally mediated communication is different to other genres of writing, I examined various features of language variation (McMenamin 2002 and Olsson 2008), and style (Coulthard 2005 and Crystal 2011), and how these informed a forensic linguistic study conducted in digitally mediated communication. This led to a discussion on the style markers I intended to use in the study, starting with the stylistic style markers: spelling, punctuation and typography; lexis; and DMC features, and then discussing the stylometric style markers: Keywords (Kotzé 2010 and Scott 2012), function words (Morton 1978 and Hubbard 1995), the most frequently occurring words, and punctuation and spelling (Chaski 2001 and Olsson 2008).

Chapter 3 dealt with the research methodology. It started by examining the ethical issues involved in using other people's texts for research purposes, and giving an account of how the participants and data were selected. It then described how the qualitative research was conducted, firstly by describing the aspects of what constitutes a stylistic analysis before looking at how the stylistic features were to be categorised. Secondly, the Scientific Working Group for Forensic Document Examination (SWGDOC) Band scales were discussed, as its scoring criteria were used to draw conclusions as to which text was the most likely to have been written by Writer X. The quantitative analysis section began by describing the different tools available on *WordSmith Tools,* namely, *KeyWords, Concordance* and *WordList,* which were used to conduct the stylometric analyses. After a discussion of the statistical test to be used, the chapter moved on to show how each of the chosen style markers, namely, keywords, function words, most frequently occurring words and punctuation were to be analysed using *WordSmith Tools.*

Chapter 4 commenced by qualitatively examining the participants' writings against the disputed text (X). Despite all the writers coming from similar sociocultural backgrounds, their individual writings exhibited some variation. The qualitative assessment showed how each writer's submission was noticeably unique, for although they all shared the same appropriately chatty style, all the writers displayed different lexical choices ranging from the quite formal (*keep abreast*) through to features usually associated with text messaging (*gr8*) (Crystal 2009). The most noticeable feature was the variation in punctuation usage, from the adherence to prescriptive rules to the use of multiple question marks and exclamation marks together with emoticons. The qualitative assessment allowed me to extract features such as idiosyncratic punctuation, which were so common that many could be measured statistically. The qualitative assessment was conducted at both the1,000-word and 2,000-word levels, and in both cases the style markers accurately identified the writer of the disputed text. The qualitative analysis was aided by the SWGDOC band scales, a version of which is used by the FBI to analyse handwritten documents. The descriptors of the SWGDOC band scale, although open to a degree of subjectivity in their interpretation, were helpful in correctly identifying Writer G as being the author of the disputed text (X).

The second part of the research was the quantitative analysis. Using the concordance program *WordSmith Tools:* Keyness, function words, the most frequently occurring words and punctuation were analysed and tested statistically, using Chi-square at both the1,000-word and 2,000 word levels. In interpreting the findings, consideration was not only given to the issue of whether they indicated that writers should be excluded because they were sufficiently different to X ($p>0.05$), but also to the extent to which the statistics indicated that a writer should be regarded as sufficiently similar to Writer X to be identified as one and the same writer. Grant and Baker's (2001) threshold of $p>0.95$ was not met in any of the stylometric analyses. It would be difficult to reach the $p>0.95$ level in forensic cases where the text lengths are relatively short. The quantitative analysis also correctly identified G as the author of the disputed text (X) in all the tests.

5.3 Contribution of the study

In this section I try to assess the contribution this study has made to authorship attribution within the field of forensic linguistics, which will be discussed in terms of its aims, looking initially at the overall aim – most specifically with respect to the focus on authorship analysis on Facebook. This is followed by consideration of the two objectives of the more specific Aim 1, which relate to the effectiveness of the stylistic and stylometric investigations respectively, and in both of these sections, discussion includes an assessment of the contribution of Aim 2, which relates specifically to the role of text length in authorship attribution.

5.3.1 Facebook language and authorship analysis

The language used on Facebook and social networking sites in general appears to be underdescribed, with the exceptions of Crystal's (2011) work on Twitter and Olsson's (2009b) study on mobile telephone texting. In my study, the examples of language use from the participants' Facebook writings and the subsequent analyses should add to a growing body of knowledge of how language is being used, and is evolving in DMC. It should add some weight to Olsson's (2009b) conclusion, in his study regarding language use in mobile telephone messaging, where he found that women were using less conservative language than men when texting. Although my study does not investigate the language of men on Facebook, it does show that the well-educated, highly literate women writers are both willing and able to use informal language, based largely on spoken registers, when they feel like it, following a trend in DMC for social purposes in general.

More importantly, this study, despite its focus on what might be called a single sub-genre of digitally mediated communication (Facebook), and also the tightly defined demographic of the selected writers in terms of home language, education level, age and gender, achieved its general aim of exploring the extent to which it is possible to attribute authorship and its findings reveal that it is indeed possible to do so, to a very considerable extent. However, the extent to which authorship can be attributed on Facebook is dependent on the relevant features being tested and the length of text.

There is considerable evidence that the methods employed in this study are effective in identifying the writer of the disputed text correctly. The three sections that follow will deal with the stylistic, stylometric and text length issues respectively.

5.3.2 Stylistic analyses

From Writer X's text, seven stylistic features were extracted and of those, two were not exhibited in the 1,000-word text, namely, *word+no space+ellipsis (4 dots)+no space+word* and *apostrophe+cause ('cause).* The other eight texts were then mined for the remaining five stylistic features. As was seen in Table 4.17 at the1,000-word level, all the writers shared some of the stylistic features with X, with A as well as G sharing all five. However, one of A's shared features showed a statistically significant difference with X, namely *word+space+dash+space+word* (which was significant in Writers C and E too). In addition to shared stylistic features, the eight texts were examined for features not shared with X. Although F and H were similar to G in not having any unshared features with significantly different frequencies to X (with of course a nil frequency), G only exhibited two features not shared with X, as opposed to four for F and nine for H. Thus in my study, by considering the number of features shared with the 'disputed' text, the number not shared, and where applicable, significant differences in frequencies of both shared and unshared features, a stylistic approach was applied that, even at the 1,000-word level arrived at a finding that indicated quite strongly that G was the most likely writer of the X text.

At the 2,000-word level, all the writers had features shared with X, but G was the only one that shared all seven features, as was seen in Table 4.18. Writers A and G shared one feature with X, namely, *capital letters for emphasis,* which showed a statistically significant difference with X. Four writers, namely, A, C, E and F exhibited the construction *word+space+dash+space+word,* and Writers A and E used *brackets* at significantly different levels to X. Writer G was also the only author not to have any unshared features exhibited at significantly different frequencies to X. Therefore, in this study, the number of shared features, in combination with the unshared features and their respective significant differences, underline the effectiveness of the stylistic

approach to identifying the author of the ‘disputed text’, which in this case was Writer G, at the 2,000-word level.

Tables 4.17 and 4.18 only showed the stylistic features which were significant, but there were many other stylistic features whose frequency counts did not reach a significant level at the 1,000-word and 2,000-word levels. Despite that, they were still included in the stylistic analysis as they gave a broader picture of the nature of the text. A further observation was that, with the exception of using the lowercase *i* for the first person pronoun *I* by D, all the shared and unshared features that exhibited significant differences were aspects of punctuation. This relates directly to my first aim of testing the extent to which the style markers are effective in identifying the author of the disputed text. Punctuation, as a style marker, correctly identified G as the most likely author of the disputed text at both the1,000-word and 2,000-word levels.

The second method used to qualitatively assess the texts was to assign band scores using the SWGDOC band scales, which are a useful tool in analysing the qualitative data and for comparing features which are relevant but not statistically expressible McMenamin (2002).

> The majority of forensic linguists and phoneticians have traditionally felt that they were unable to express their findings statistically in terms of mathematically calculated probabilities and so have expressed them as semantically encoded opinion. (Coulthard 2010, 480)

5.3.3 Stylometric analyses

The stylometric analysis was made up of four tests, namely, keywords, function words, most frequently occurring words and punctuation. At the 1,000-word level, the keyword test correctly identified G as the most likely author of the disputed text, as she had only two keywords and an average keyness value of 4.07. However, the fact that E had three keywords and an average keyness value of 4.54 put her fairly close to G. This highlights the fact that keywords need to be used with a measure of caution when one is analysing

a relatively short text. At the 2,000-word level, G was again correctly identified as the most likely author of the 'disputed' text by a convincing margin. G only had seven keywords and an average keyness of 5.25 (Table 4.10). The next closest to G was H with 13 keywords. Despite E having the same average keyness as G (5.25), she had 15 keywords (Table 4.10), which effectively excluded her from being the author of the 'disputed' text. It is difficult to assess the effectiveness of keywords as a style marker in comparison to the other three tests, but easier to compare the latter with one another as they generate overall *p* values for each writer at each text length.

The 23 function words recommended by Morton (1978) were the least successful of the style markers at the1,000-word level as four of the eight writers met the $p>0.05$ criterion (Table 4.19). Despite that, the function word test did identify G as the most likely author of the 'disputed' text with a *p* value of 0.76, but, with writers E, B and D exhibiting *p* values of 0.45, 0.17 and 0.14 respectively, there is an element of doubt as to how effective this style marker is. The same style marker at the 2,000-word level showed a very different picture. Despite G exhibiting a relatively low *p* value of 0.16, she was the only writer to reach the $p>0.05$ criterion (Table 4.20).

The most frequently occurring words at the1,000-word level showed somewhat more promise than the function words. Again, G was correctly identified as the most probable author of the 'disputed 'text' with a *p* value of 0.86 (Table 4.19), and the only other writer to reach the $p>0.05$ criterion was E with a *p* value of 0.48. At the 2,000-word level, only G exceeded the $p>0.05$ level at a convincing 0.88 (Table 4.20). It could be argued that using the most frequently occurring words is a better prospect than using Morton's (1978) pre-determined list as the words come directly from the participants and thereby remove the need to exclude certain words due to genre influence as was experienced by Hubbard (1995).

The punctuation style marker was the most successful of the three. Despite G only attaining a *p* value of 0.46 at the 1,000-word level, she was the only author to reach the $p>0.05$ criterion (Table 4.19). This trend was repeated at the 2,000-word level, where G

exhibited a *p* value of 0.60. A possible reason for the success of the punctuation style marker is the nature of punctuation usage in DMC, which is characterised by far greater freedom for idiosyncratic use, and the fact there is more likelihood of having more countable features (Olsson, 2008). This means that even at 1,000 words, there would likely be enough countable features to carry out a statistical analysis and there would likely be sufficient variation between the writers due to idiosyncratic usage. However, despite these promising results, punctuation should be treated with a degree of caution. Would these same results occur if the writing had used a medium less tolerant of punctuation creativity, such as a work-related e-mail? Moreover, punctuation creativity is very much under the conscious control of the writer and is arguably the easiest feature to fake. All of my participants are well educated and are employed in professional occupations, it follows that they are well versed in standard punctuation usage, and yet they chose to use non-standard forms on Facebook.

The stylometric analyses of the three style markers all correctly identified Writer G as the author of the 'disputed' text, and although none of her *p* values reach the 0.95 level (Grant and Baker 2001), they are, nevertheless, still effective at attributing authorship. The order of effectiveness with reference to *p* values is as follows: (1) most frequently occurring words (2,000); (2) most frequently occurring words (1,000); (3) function words (1,000); (4) punctuation (2,000); (5) punctuation (1,000) and (6) function words (2,000). However, even more important than the absolute *p* values for G are the *p* values for the other writers in comparison. Thus, function words (1,000) shows three other writers above 0.05, with E at 0.45 and the most frequently occurring words (1,000) has one at 0.05 and another at 0.48 and only in the remaining four groups are all the other writers on a *p* value less than 0.05. When values for competing writers are also considered, the final order of effectiveness is (1) most frequently occurring words (2,000); (2) punctuation (2,000); (3) punctuation (1,000) and (4) function words (2,000). Function words then seem to be the least effective of the three.

The question of text length and the application of stylometry within a forensic linguistic context on Facebook, or any form of digitally mediated communication is, arguably, the

most relevant aspect that needs to be considered. An excellent real world example is the Zuckerberg/Ceglia case investigated by McMenamin (2011). The frequency counts McMenamin reached were too small to be statistically analysed, but he was still able to reach a conclusion based on a battery of observable style markers. Proponents of quantitative research were swift to condemn McMenamin's conclusions as being unscientific as he did not assign any statistical value to his findings. Chaski (2001, 2) referred to such methods as "junk science". The problem is that most texts originating from DMC such as Facebook postings, tweets, text messages and e-mail are likely to be "unhelpfully short" (McMenamin 2002, 181) and "most suicide notes and threatening letters, for example, are well under 200 words long, and many consist of fewer than 100 words" (Coulthard 2004, 2). In my simulated study, my participants had substantial amounts of writing in their Facebook communications, but in the real world that may not always be the case. Does that mean we have to abandon any authorship attribution attempts if there is not a sufficient word count? One of the implications of my study is that a combination of stylistic and stylometric approaches, both carefully considered, should be used together as far as possible, although this will not be feasible when texts are particularly short.

5.4 Limitations of the study

This section deals with some of my study's limitations including with respect to style markers and the data collected and comparisons made.

5.4.1 Style markers

McMenamin (2002) listed over 300 style markers that had been used in over 80 authorship attribution cases, some of which were used in my study. Even though the style markers chosen for the stylistic analysis were informed by Writer X's text, there were a few features that were either minimally acknowledged or ignored completely, and these could also have been analysed. Table 5.1 gives an overview of some of these.

Table 5.1 Potential style markers not analysed

Stylistic feature	Comment
Greetings	Writer X began most of her threads with the greeting *Hi+name+comma*
Salutations	Writer X frequently ended her messages with *Lots of love*
Sentence length	Writer X's sentences tend to be short with minimal use of conjunctions, relative clauses or other linking devices.
Paragraphing	Writer X tends to follow the prescriptive letter writing paragraphing styles.
Repetition of words	Writer X had some examples of words being repeated three times: *fun fun fun* and *welcome welcome welcome*

Another issue which was not taken into account was that of grammar. Although the participants had few errors, there were a number of idiosyncratic uses, for example, the omission of the personal pronoun in *would love to hear.* It is the norm on Facebook to use a more informal spoken register, and this was the case, in varying degrees, with all eight writers. Even though many aspects of Facebook discourse were covered in the analyses, it was still beyond the scope of this study to undertake a systematic account of this type of language.

5.4.2 Data collected and comparisons made

My study focused exclusively on comparing Writer X with the other writers, but what was not addressed, was how different the other writers were from each other. If two writers whose writings appear to be similar were to be compared using the various stylistic and stylometric analyses, would there be significant idiolectal difference? Tests such as that could help strengthen or weaken the case for idiolect in authorship attribution. The data collected for my study was from a tightly defined group of writers in terms of home language, education, age and gender. As there was not any data from male writers, this meant that I could not investigate gender as a variable in this context and could not follow up on Olsson's (2009b) research, where he found that men were more conservative writers in mobile phone texting, a form of digitally mediated communication. If male writers had been included, it may then have been found that with fewer idiosyncrasies, there would have been fewer differences and statistically less

significant ones for certain style markers, and so the effectiveness of some of them in terms of authorship would have been lower.

A further comparison that could have made was to compare each writer's first 1,000 words with their second 1,000 words to explore intra-author consistency. Moreover, each writer's second 1,000 words could have been compared to those of X. However, due to the limitations of space and time, it was decided that this was beyond the scope of the present study.

5.5 Suggestions for further research

Given the relatively narrow focus of my study, both in terms of my chosen demographic and my text, there is a great deal more research that needs to be done if we are to fully understand how stylistics and stylometry in the domain of authorship attribution apply to Facebook and social networking sites.

5.5.1 Text analysis of male authors

As discussed also in the previous section my research focussed exclusively on females within a certain demographic, and the question to ask is: would males in the same demographic be any different? There has been a great deal of sociolinguistic work conducted on the differences in language usage between males and females (Trudgill 2001), and Olsson's (2009b) findings that men are more conservative language users in digitally mediated communication is in contrast to the standard orthodoxy in sociolinguistics that women use more prestige forms than men (Trudgill 2001).

5.5.2 Second language speakers

According to the social media analyst company *Social bakers* (www.socialbakers.com), the country with the third highest number of Facebook users is India, with just over 61 million subscribers and in eighth place is the Philippines, with just over 30 million subscribers. These two countries are both in Kachru's (1997) expanding circle of countries which use English as a second language. This highlights the potential for authorship attribution research among English second language (and foreign language)

speakers. It is interesting to note in this connection the report in The Mail Online on how Al Qaeda is using Facebook for recruitment purposes through their English language Facebook page (Gardner 2010).

5.5.3 Facebook status updates and group threads

The texts I used for my study came from the Facebook inboxes of my participants, which is only one of three modes of communication on Facebook, the other two being status updates and instant messaging. Similar to status updates are postings on the discussion threads of the various groups. Figure 5.1 shows an example of a discussion thread from a controversial Facebook group called Global Secular Humanist Movement. It is in threads like this that extremists have been known to threaten people while hiding behind a false identity or avatar. However, unlike Facebook inboxes, such threads are unlikely to consist of even1,000 words of text from a single person. The implications of this are that the researcher would more likely have to rely on stylistic analyses alone rather than in combination with stylometric analyses.

Figure 5.1 – Facebook group thread

5.6 Conclusion

With the advent of social networking sites and Facebook in particular, a whole new world of possibilities has opened up for forensic linguistic research. This is especially so if one considers that Facebook is nearing one billion subscribers and has become a political force for social change, with politicians using it for campaigning, Al Qaeda using it to spread its message and disaffected youths in London using it to organise riots. It is my hope that this study has made some contribution to the field of authorship analysis in general and how it applies to social networking sites such as Facebook in particular. The success achieved in this study in exploring the effectiveness of author attribution derives partially from the fact that both stylistic and stylometric analyses were used, and I hope that the study can be seen as exemplifying the following point:

> *I believe that the application of the combined approach to a wider range of text types should lead to an increasing refinement of this methodology in future.*
>
> (Kotzé, 2010: 195)

References

Acquisiti, A. and Gross, R. 2006. *Imagined communities: awareness, information sharing and privacy on the Facebook*. Cambridge: Robinson College.

Baayen, H. Tweedie, F.J, Neijt, A.H, Halteren, H and van Krebbers, L. 2000. Back to the cave of shadows: stylistic fingerprints in authorship attribution. The ALLC/ACH2 000 Conference. University of Glasgow (Unpublished paper).

Baron, N. 2003. Why email looks like speech: proofreading pedagogy and public face. In *New Media Language*, ed. J. Aitchison and D. M. Lewis, 85–94. London: Routledge.

Becker, J. A. H. and Stamp, G. H., 2005. Impression management in chat rooms: A grounded theory model. *Communication Studies* 56:243 -260.

Bednarek, M. 2009. Corpora and discourse: a three-pronged approach to analysing linguistic data. In *Selected proceedings of the 2008 HCSNet workshop on designing the Australian National Corpus,* ed. M. Haugh, K. Burridge, J. Mulder and P. Peters,19-24. Somerville, MA: Cascadilla Proceedings Project.

boyd, d. m. and Ellison, N. B. 2007. Social networking sites: definition, history and scholarship. *Journal of Computer-mediated Communications* 13:210 -230. http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html (accessed 14 March 2011).

boyd, d. m. and Hargittai, E. 2010. Facebook privacy settings: who cares? *First Monday* 15 (8). http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3086 (Accessed 14 March 2011).

Burnard, L. 1995. *User's reference guide for the British National Corpus*. Oxford: Oxford University Computing Services. http://homepages.abdn.ac.uk/k.vdeemter/pages/teaching/NLP/practicals/bnc-doc.pdf (Accessed 9 December 2012).

Burrows, J. F. 1992. Computers and the study of literature. In *Computers and written text: an applied perspective*, ed. C. Butter, 167-204. Oxford: Blackwell.

Chaski, C. E. 2001. Empirical evaluations of language-based author identification techniques. *Forensic Linguistics* 8(1):1-65.

Chaski, C. E. 2005. Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence,* 4(1):1-13.

Chaski, C. E. 2011. High-stakes forensic linguistics [Online forum]. http://languagelog.ldc.upenn.edu/nll/?p=3309 (Accessed 20 May 2012).

Cochran, W. G. 1954. Some methods for strengthening the common (chi-square) tests. *Biometrics* 10:417–451.

Cohen, L. J. 1977. *The probable and the provable.* Oxford: Clarendon.

Coulthard, M. 2000. Whose text is it? On the linguistic investigation of authorship. In *Discourse and Social Life*, ed. M. Coulthard and S. Sarangi, 271-287. London: Longman.

Coulthard, M. 2004. Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics* 25(4):431-447.

Coulthard, M. 2005. *Some forensic applications of descriptive linguistics.* http://www.ufjf.br/revistaveredas/files/2009/12/artigo016.pdf (accessed 22 June 2013)

Coulthard, M. 2010. In my opinion. In: *Routledge handbook of forensic linguistics,* ed. M. Coulthard and A. Johnson, 508-522. New York: Routledge.

Coulthard, M and Johnson, A. 2007. *An introduction to forensic linguistics.* London: Routledge.

Coulthard, M. Grant, T and Kredens, K. 2011. Forensic linguistics. In *The SAGE handbook of sociolinguistics,* ed. R. Wodak, B. Johnstone and P. E. Kerswill, 529-544. London: SAGE Publications.

Crystal, D. 2001. *Language and the Internet*. Cambridge: CUP.

Crystal, D. 2007. *Language and the Internet.* Cambridge: CUP.

Crystal, D. 2009. *Txtng: The gr8db8.* Oxford: Oxford University Press

Crystal, D. 2011. *Internet linguistics: a student guide.* New York: Routledge.

Crystal, D. and Davy, D.1969. *Investigating English style.* London: Longman.

Cyberbullying Research Center. 2010. …identifying the causes and consequences of online harassment. [Internet] 14 June. http://www.cyberbullying.us/research.php. (accessed 14 June 2010).

Danet, B. 2001. *Cyberpl@y: communicating online.* New York: Berg.

Davis, B. and Brewer, J. 1997. *Electronic discourse: linguistic individuals in virtual space.* Albany, NY: State University of New York Press.

Donath, J., Karahalios, K. and Viégas, F. 1999. Visualizing conversation. *Journal of Computer-mediated Communication* 4(4). [Online Version]. http://jcmc.huji.ac.il/vol4/issue4/donath.html (accessed 10 February 2013).

Gabrielatos, C. and Marchi, A. 2011. Keyness: Matching metrics to definitions. Paper presented at *Theoretical-methodological challenges in corpus approaches to discourse studies: and some ways of addressing them.* Portsmouth. Powerpoint presentation. http://eprints.lancs.ac.uk/51449/4/Gabrielatos_Marchi_Keyness.pdf (accessed 5 November 2012).

Goutsos, D. 1995. Review article: forensic stylistics. *Journal of Forensic Linguistics*. 2(1):99-113.

Grant, T. 2007. Quantifying evidence in forensic authorship analysis. *International Journal of Speech Language and the Law* 14(1):1-25.

Grant, T. 2010. Txt 4n6: idiolect free authorship analysis? IN: *Routledge handbook of forensic linguistics,* ed. M. Coulthard and A. Johnson, 508-522. New York: Routledge.

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

Grieve, J. W. 2005. Quantitative authorship attribution: A history and an evaluation of techniques. MA thesis. Simon Fraser University. https://perswww.kuleuven.be/~u0064311/MAThesis.pdf (accessed 12 May 2012).

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

Halliday, M. A. K. 1989. *Spoken and written language,* 2nd Edition. Oxford: OUP.

Hancock, I. 1986. The cryptolectal speech of the American roads: traveller cant and American Angloromani. *American Speech,* 61(3): 206-226. http://www.radoc.net/radoc.php?doc=art_c_language_traveller_talk&lang=en&articles=true (accessed 17 June 2013)

Haythornthwaithe, C. 2005. Social networks and Internet connectivity effects. *Information, Communication & Society* 8(2):125-147.

Herring, S. 2007. A faceted classification scheme for computer-mediated discourse. Indiana University: Bloomington. http://www.languageatinternet.org/articles/2007/761/Faceted_Classification_Scheme_for_CMD.pdf (accessed 22 June 2013).

Hinduja, S. and Patchin, J. W. 2009. *Bullying beyond the schoolyard: preventing and responding to cyberbullying.* Thousand Oaks, CA: Corwin Press.

Holmes, D. 1985. The analysis of literary style – A review. *The Journal of the Royal Statistical Society* 148:328-341.

Holmes, D. 1994. Authorship attribution. *Computers and the Humanities 28:*87-106

Holmes, D. 1997. Stylometry: its origins, development and aspirations in the state of authorship attribution studies: (1) the history and the scope; (2) the problems – towards credibility and validity. http://opim.wharton.upenn.edu/~sok/papers/r/s004.html. (accessed 3 December 2011).

Holmes, D. 1998. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing* 13:111-117.

Hubbard, E. H. 1995. Linguistic fingerprinting? A case study in forensic stylometrics. *South African Journal of Linguistics*: *Supplement* 26:55-72.

Instat Plus 2005. Version 3,036. Reading: Statistical Services Centre, University of Reading.

Johnstone, B.2000. *Qualitative methods in sociolinguistics.* New York: Oxford University Press.

Juola, P. 2006. Authorship attribution. *Foundations and Trends® in Information Retrieval* 1(3):233-334.

Kachru, B. B. 1997. World Englishes and English-using communities. *Annual Review of Applied Linguistics* 17:66-87

Katzman, S. L. and Witmer, D. F. 1997. Online smiles: does gender make a difference in the use of graphic accents? *Journal of Computer-Mediated Communication* 2(4) http://jcmc.indiana.edu/vol2/issue4/witmer1.html (accessed 10 February 2012).

Kirkpatrick, D. 2010. *The Facebook effect: The inside story of the company that is connecting the world.* New York: Simon and Schuster.

Kotzé, E. F. 2010. Author identification from opposing perspectives in forensic linguistics. *Southern African Linguistics and Applied Language Studies* 28(2):185-197. http://www.ajol.info/index.php/salas/article/view/59786. (accessed 5 November 2012).

Labov, W. 1994. *Principles of linguistic change. Volume 1: Internal factors*. Oxford: Blackwell Publishers.

Leech, G. and Smith, N. 2005. Extending the possibilities of corpus based research on English in the twentieth century: a prequel to LOB and FLOB. *ICAME Journal* 29:83-98.

Leonard, R. A. 2005. Forensic linguistics: applying the scientific principles of language analysis to issues of the law. *The International Journal of the Humanities* Vol 3:1-10. http://www.robertleonardassociates.com/PDF/ForensicLinguistics_Applying-Scientific-Principles.pdf. (accessed 10 February 2012)

Liberman, M. 2011. High-stakes forensic linguistics [Internet]. http://languagelog.ldc.upenn.edu/nll/?p=3309. (accessed 20 May 2012).

Madden, M. and Smith, A. 2010. Reputation management and social media. *Pew Internet & American Life Project.* http://www.pewinternet.org/Reports/2010/Reputation-Management.aspx. (accessed 4 April 2011).

Martine, R. 2012. Group Facebook. Improve.com [Internet]. http://www.improve.com/discussion/id_group/1822/id_group_post/1843. (accessed 10 November 2012).

Matthews, P. H. 1997. *Oxford concise dictionary of linguistics.* New York: Oxford University Press.

McMenamin, G. R. 2002. *Forensic linguistics: advances in forensic stylistics.* London: CRC Press.

McMenamin, G. R. 2010. Forensic stylistics: theory and practice of forensic stylistics. In: *Routledge Handbook of Forensic Linguistics*, ed M. Coulthard and A. Johnson, 487-507. New York: Routledge.

McMenamin, G. R. 2011. Style markers in QUESTIONED vis-à-vis KNOWN - Zuckerberg. http://www.scribd.com/doc/56976903/W-D-N-Y-1960merged-Linguist (accessed 21 June 2013)

Ceglia v. Zuckerberg, Case 1: 10-cv-00569-RJA-LGF, *United States District Court Western District of New York,* 2011.

McNaughton, M. 2012. Social networking stats: Facebook to reach one billion users by August. *The Realtime Report,* [Internet] 13 January. http://therealtimereport.com/2012/01/13/social-networking-stats-facebook-to-reach-one-billion-users-by-august-rltm-scoreboard/ (accessed 7 June 2012)

Messmer, E. 2007. Study: Facebook users easy targets for identity theft. NetworkWorld.com [Internet] 14 August. http://www.networkworld.com/news/2007/081407-facebook-identity-theft.html. (accessed 1 September 2011).

Millard, W.B. 1996. 'I flamed Freud: A Case Study in teletextual incendiarism'. In *Internet culture*, ed. D. Porter, 145-159. New York: Routledge.

Morton, A. Q. 1978. *Literary detection: how to prove authorship and fraud in literature and documents.* Bath: Pittman Press.

Nickson, C. 2009. The history of social networking. [Internet] 29 January 2009. http://www.digitaltrends.com/features/the-history-of-social-networking/. (accessed 1 December 2010).

Olsson, J. 2004. *Forensic linguistics.* London: Continuum International Publishing Group.

Olsson, J. 2008. *Forensic linguistics.* 2nd Edition. London: Continuum International Publishing Group.

Olsson, J. 2009a. *Word crime: solving crime through forensic linguistics.* London: Continuum Publishers.

Olsson, J. 2009b. Preliminary observations on author variation in mobile phone texting. The Forensic Linguistics Institute in association with the Forensic Linguistic Society. (Unpublished).

Olsson, J. 2010. The Facebook killer, [YouTube]. http://www.youtube.com/watch?v=Dy4fYa-NZPk (accessed 26 October 2010).

Raising children network. 2006. Protecting children from paedophiles, [Internet] (Updated 08 August 2006) http://raisingchildren.net.au/articles/protect_your_child_from_pedophiles_-_cyh.html (accessed 09 February 2012).

Rudman, J. 1998. The state of authorship attribution studies: some problems and solutions. *Computers and the Humanities* 31(4):351-365.

Sapir, E. 1949. *Language*. New York: Harcourt, Brace and Company.

Scott, M. 2012., *WordSmith Tools Manual, version 6*. Liverpool: Lexical Analysis Software Ltd.

Shapero, J. J. 2011. The language of suicide notes. PhD thesis, University of Birmingham, URL http://etheses.bham.ac.uk/1525/. (accessed 13 June 2013).

Stanford, A. J., Aked, J. P., Moxey, L. M and Mullen, J. 1994. A critical examination of assumptions underlying the CUSUM technique of forensic linguistics. *International Journal of Speech Language and the Law.* 1(3):151-167

Sundén, J. 2003. *Material virtualities.* New York: Peter Lang.

Svartvik, J. 1968. The Evans statements: A case for forensic linguistics. Stockholm: Almqvist and Wiksel.

Swales, J. 1990. The concept of the discourse community. Boston: Cambridge University Press.

Trudgill, P. 2001. *Sociolinguistics: an introduction to language and society.* 4th ed Penguin books.

Walkely, A. 2009. Typing friendships into being: vocatives in Facebook wall-to-wall conversations. University of Sydney.

Popular media sources (online and paper based newspapers, and magazines)

Beaver, T. 2008. Bullies slip through the Net. *The Star,* 29 April.

Carter, H. 2010. Facebook killer sentenced to life for teenager's murder. *The Guardian*, [Internet] 8 March. http://www.guardian.co.uk/uk/2010/mar/08/peter-chapman-facebook-killer (Accessed 7 June 2010).

FOXNews. 2009. Teen accused of blackmailing classmates for sex acts in Facebook scam. *FOXNews.com,* [Internet] 6 February. http://www.foxnews.com/story/0,2933,488428,00.html (accessed 7 June 2010).

Gardner, G. 2010. Facebook used by Al Qaeda to recruit terrorists and swap bomb recipes, says US homeland security report. *Dailymail Online* [Internet]. http://www.dailymail.co.uk/news/article-1337344/Facebook-used-Al-Qaeda-recruit-terrorists-swap-bomb-recipes.html. (accessed 21 January 2013).

McDonald, S. N. 2005. Facebooking, the rage on college campuses. *The Seattle Times* [Internet] 4 July. http://community.seattletimes.nwsource.com/archive/?date=20050704&slug=btfacebook04 (accessed 22 June 2011).

MSN BC. 2009. Facebook ID theft targets 'friends'. [Internet] 30 January. http://redtape.msnbc.msn.com/_news/2009/01/30/6345792-facebook-id-theft-targets-friends (accessed 1 September 2011).

News24.com. 2011. 'Facebook rapist' appears in Cape court, [Internet] 3 November. http://www.news24.com/SouthAfrica/News/Facebook-rapist-appears-in-Cape-court-20111103 (accessed 10 November 2011).

Salkeld, L. 2009. Facebook bully jailed: death threat girl, 18, is first person put behind bars for vicious Internet campaign. *The MailOnline.* http://www.dailymail.co.uk/news/article-1208147/First-cyberbully-jailed-Facebook-death-threats.html (accessed 3 November 2011).

Savill, R. 2009. 'Tweetups' and 'unfriend' among Oxford English Dictionary's 'words of the year'. *The Telegraph,* [Internet] 17 March. http://www.telegraph.co.uk/news/newstopics/howaboutthat/6905776/Tweetups-and-unfriend-among-Oxford-English-Dictionarys-words-of-the-year.html. (accessed 15 March 2011).

Seamark, M. 2010. Paedophile postman used Facebook and Bebo to groom up to 1,000 children for sex. *MailOnline,* [Internet] 28 May. http://www.dailymail.co.uk/news/article-1282157/Facebook-grooming-How-pervert-postman-used-site-groom-hundreds-children.html (accessed 14 June 2010).

Sengupta, S. 2012. Facebook's prospects may rest on trove of data. *New York Times,* [Internet] 14 May. http://www.nytimes.com/2012/05/15/technology/facebook-needs-to-turn-data-trove-into-investor-gold.html?_r=1 (accessed 15 June 2012).

Sullivan, M. 2007. Is Facebook the new MySpace? *PC World.* http://www.pcworld.com/article/134635/is_facebook_the_new_myspace.html. (accessed 21 June 2011).

Taylor, J. 2009. The Facebook paedophile ring. *The Independent,* [Internet] 2 October. http://www.independent.co.uk/news/uk/crime/the-facebook-paedophile-ring-1796373.html (accessed 07 June 2010).

Whittaker, L. 2010. Facebook speak: Teenagers create secret online language. *The Telegraph,* [Internet] 26 April. http://www.telegraph.co.uk/technology/facebook/7632133/Facebook-speak-Teenagers-create-secret-online-language.html (accessed 20 August 2011).

Zimmer, B. 2011. Decoding your e-mail personality. *The New York Times,* [Internet] 23 July. http://www.nytimes.com/2011/07/24/opinion/sunday/24gray.html (accessed 16 July 2011).

Cases cited

Ceglia v. Zuckerberg, Case 1: 10-cv-00569-RJA-LGF, *United States District Court Western District of New York,* 2011.

England & Wales v Cartwright [2010] T20097610.

SCSA (Supreme Court of South Africa). 2004. Judgement in the matter between the State and Luzuko Kerr Hoho. Case CC 132/02, pp 2814-2896 (Judge CS White).

Appendix 1

Permission letter

PERMISSION TO CONDUCT RESEARCH

I am an MA student in the Linguistics department of the University of South Africa (UNISA) and I am currently conducting stylometric research on the language used on social networking sites, specifically Facebook.

It would contribute greatly to my research if I could make use of writing samples from your Facebook inboxes.

Please note that all participation is voluntary and you may withdraw at any time. Moreover, all names used in any written and spoken reporting of my research will be removed and replaced with pseudonyms, thereby ensuring your privacy and anonymity.

Should you require any further details, you are welcome to contact either me or my supervisor, whose details appear below.

Sincerely

Colin Michell
4322-607-9
Tel: 00971 50 837 2217
E-mail: cmichell@hct.ac.ae

Supervisor: Prof. Hilton Hubbard
Tel: 012 429 6732
e-mail: hubbaeh@unisa.ac.za

I …… grant Colin Michell permission to make use of writing samples from my Facebook inbox.

…… ……
(signature) *(date)*

Appendix 2

Request letter

I am currently busy with my Masters Degree in Linguistics with a specialisation in Forensic Linguistics and I am doing a study on the language used on Facebook / MySpace/ Twitter and any other social networking sites. I am looking at ways of determining authorship between texts. This is where you can help – I need to collect texts of just over2 000 words from at least 16 people (8 men, 8 women). It is going to be a purely statistical study (counting how many times a person uses the word 'but' for example) so the content of the texts is completely irrelevant. I am basically trying to see which writing features (spelling, punctuation, slang words etc) can be used to separate authors on social networking sites.

The demographics for my study are: males and females aged between 30 and 40 who went to an English medium school in South Africa. The idea behind such a tight demographic is that they should have a similar style of writing and it should be more difficult isolate the different authors.

If you would like to take part in the study – I need from you:
+/-2 000 words from your facebook messages – just your writing not the responses, cut and pasted onto a word document and e-mailed to

colin@ihjohannesburg.co.za

For the purposes of this study you will be completely anonymous, your name will not appear anywhere – you will just be a pseudonym. If you would like to see the results of the study before I present them to the good professor, that wouldn't be a problem.

The results of the study will be more tools to help catch paedophiles, 419 scammers, online bullies and anyone else who uses social networking sites like facebook to hurt people.

If you would like to discuss this study with me, please feel free to phone me on

072 224 6997 (South Africa). 0027722246997

Cheers
Colin

Appendix 3

Participants' texts

Writer X

Hi *Name*,

Happy Happy Belated birthday and SOOOO sorry I missed it. I'm a terrible friend I know. Please forgive me!

How have you been? Saw your sexy tango pics. Very cool! Are you still dancing a few times a week? Looks like fun, fun, fun.

I'm at home today trying to rest my big 'cangle' that I sprained. *Name* and *Name* are at The Tennis Club swimming so for once it's nice and quiet round these parts. Been working hard and trying to get annoying students under control and have finally managed to do that by basically being a complete 'bitch' to them. I don't smile, talk in a low, angry voice and we have NO fun in Miss *Name*'s class and I mean NO fun! It really works but it's a lot of work for me trying to maintain this personal.

How's *Name* and *Name*? What's news from your side? What's happening with '*Name*' - is that his name?

Please write when you get the chance. Would love to hear from you.

Lots of love

Hi *Name*,

Thanks for your message. I've hardly had time to do anything the last few weeks just because I've been trying to find my way round at work etc. It's going fine but still quite up and down. Hopefully this is normal and will settle down soon.

How long have you been in HK? Years now I'm sure. What do you love about it? What do you hate about it? Would love to hear more about your experience of being in a foreign country. Is your family still in SA?

You need to contact *Name Name* at the Lab. Her email address is *e-mail address*

Hope this helps,
Take care,

Hi *Name*,
Just been looking at your website and I'm like 'WOW' you're so good! Did you only

recently discover this talent of yours?

Glad you can come to the picnic. See you soon.

Hi *Name*,

Loved your pics of Colombia. What's *Name* like? Where's she from? Sorry to hear about your German guy encounter. He sounds like a complete ass and more like he lead you on than the other way round. If he's so in love with the Colombian girl then why did he stay at your place, get drunk etc. etc. ?

We're doing better over here now that we've settled in a bit. Are members of the Tennis and Country Club and feel like complete colonial snobs going there which is great! It's really nice - with pools for *Name* and nice gym etc. for us and all set next to the Hajar mountains. This evening we met *Name* for coffee at The Hilton which is another fancy hotel at the sea. *Name*'s from SA and lives with her husband *Name* in our building. She's great. Actually everyone's really nice and has moved around a lot so know how it feels.
The students are taking a while to get used to. My one class is fab and sweet but the other are complete brats. They're really low level - don't have a good foundation in the language at all. For eg. don't even know what a verb is. The public schooling system here sucks it seems. They actually write exams in English at school but can't do very well judging by their spelling and reading ability.

Think the main reason why singletons like you wouldn't like it here is that there aren't many opportunities or people to hook up with - as in meet a life partner here. It's quite a quiet and relaxed existence out here which is what I wanted.
Have you heard from your mom and sis yet? Hope they are OK?

Miss you lots. There's no-one as fun to chat with at work as you. Hope things are less tense with German brat! Don't worry about *Name* and his stupid family - he's complex and you wouldn't have been happy with him ultimately. He really made you feel crap a lot of the time and his poor son is probably on a diet already!

Lots of love

Hi *Name*,
It's the weekend here already and I'm sitting at home with big cankle trying to relax and not walk around too much. *Name* and *Name* have gone swimming at The Tennis Club so I've finally got the chance to catch up with you.

Where is your pub quiz going to be held? Have you been before? They have weekly quizzes here too that a lot of the HCT staff go to. Sounds like you're really getting into your life in Colombia - having dinner parties, going paragliding and now the hike! Sounds great! Do you have the week off next week? How much leave do you get?

I almost quit last week as my one class of 22 girls were COMPLETELY out of hand. I thought I was going to go crazy. So this week I changed everything about me to survive with them for the rest of the semester (another 10 weeks). I didn't smile, I spoke in a low, angry voice and I sent them to Ms *Name* (head of st services who is very scary looking and very Lesbian looking!) when they spoke back at me. It seems to be working as they are behaving like little angels now but I have to prepare myself mentally, emotionally and spiritually before going in to class now. There is NO fun in Miss *Name's* class these days and I mean NO fun. They behave like 12 year olds and don't know how to be students. It's more like teaching junior high. I'm feeling much better now that I've got them under control as it was really making me miserable.
The staff at HCT are WONDERFUL! And the lifestyle here is WONDERFUL for families. We're having a picnic next Sat for *Name*'s birthday and have invited about 30 people from work. We're going to a beautiful beach called Kalba.

_____

Any suggestions for the picnic - party treats / games etc would be welcome as I know how good you are at arranging such things.

Heard from a few people that things have been a bit tense at The Lab. Sounds like *Name* has been a bit of a tyrant and not acknowledging *Name* at all. The new ADOS sounds super organised and cute. Did you meet her? What were your last two weeks there like? Are you still glad you moved? What are your plans for the summer? How's your sister and *Name*?

I'm on Skype under *Skype address*. Are you on? Get yourself a webcam girl. We can even just chat without the camera if you don't have one yet. It really helps.

Missing you lots and lots and hope you are happy and find the perfect man SOOON! What about Name? Does he have any potential?

Lots of love

Hi *Name*,
Thanks for your message and sorry it's taken me so long to reply. I know what it's like and it's very annoying when life gets in the way of facebook time isn't it?!

You lucky thing going to England at the end of the month. Who is your gorgeous friend in Newcastle?

The past few weeks have been tough for me with difficult students. They are EXTREMELY difficult and hard to control in class. I have now become a really strict, we're-never-going-to-have-fun-again kind of teacher because that's the only thing that works. *Name* says it's good 'cause it'll help me deal with *Name* when she's a teenager. Actually it's helping me now and she's only two so this tells you something about the way the students behave (like 2 year olds a lot of the time).

Otherwise I'm really happy here. The staff at the college are FABULOUS - everyone is

so cool and interesting and wonderful. This brings me back to the subject of *Name*. Could he send you a message on facebook or should I rather not encourage him? Don't feel obliged or anything....really.

Please say hi to everyone at the Lab,
Take care
lots of love

Hi *Name*
Thanks for writing and thinking of me....I've also been terrible about being in contact with friends. Tell me more about your business - doing what? full-time? Sounds very adventurous....well done!!

We're so busy doing last minute packing, seeing people for the last time...it's quite exhausting. I hope you will make it down here as I'd love to see you but I really understand if it won't be possible. We're leaving on the 8th Jan and you guys are welcome to visit us over there ANY TIME!

How are the boys? *Name* had another boy 4 weeks ago called *Name*....he's gorgeous. It's going to be hard to leave them all but I'm trying to be strong.

At the moment I'm writing from my parents' computer as both of ours are in for repairs so don't think the skype thing is going to happen anytime soon. Once we're in the Emirates we'll definitely be better set up.

Love to your family and have a wonderful Christmas.
Love

Hi *Name*

Thanks for you mail. Finally someone who knows how to pronounce '*Name*' which is the pron of '*Name*'. We thought that everyone would call her *Name* if we spelt it like that so made it more of an Indian spelling but people still get it wrong. What can you do?!!

Some exciting news....we've just heard that they want to offer us a job in Fujairah at HCT. You're in Al Ain right? We had our interviews last week and had heard a lot of good things about Fujairah so were thrilled to hear this news. Any advice you can offer us would be welcome, welcome, welcome. How is life there? What are the pros and cons of living there and working for HCT? Do you know anything about baby care by any chance? What questions should I ask about my contract? How does the salary compare to The Lab's salary and general cost of living? Would we be able to survive on one salary only should I need to stay home with *Name*?

I'm SO excited. We've been at The Lab for 10 years now and it's really time for a change but please don't say anything about this just yet on the public section of facebook as I only want to tell The Lab once we've finalised our contracts etc.

Great to hear you're doing your masters! I'd love to hear more about it - what uni are you doing it through? What are you planning to do for your dissertation? I did things backwards i.e. completed my masters last year and am now in the middle of my DELTA. I've finished module 1 and would like to look into the other modules next year. Do they offer them at HCT?

Everything's pretty much the same at The Lab, although there are a few new faces. I was acting DoS for 1 1/2 years while Heather set up our IELTS Testing Centre. I loved doing the DoS work and combining teaching and admin. Is there any scope for this over there?

Are CELTA's run at the colleges? What are the students like? Sorry about all these questions but I figured you'd be the best person to ask.

Please post your wedding pics soon...would love to see them. Sounds very romantic indeed!

Hi *Name*,

Thanks so much for your response. It really helps to have some first hand info. about the place. I'm sure it's going to take a while to adjust but they sound very geared towards ex-pats over there so hopefully we won't be too homesick.

It will be great not to have to worry about money all the time. We're still waiting for our contracts so not sure yet exactly what we'll be earning. Hope it comes through soon as we really don't want to leave the Lab in the lurch and would like to tell them sooner rather than later.

Writer A

Would love to help, only problem is I have not been on FaceBook long enough to create1 000 words. Let me know if Encouraging you all to write what your funniest or most awful experience was in the school (please dont use expletives) a Thanks for the day at the Festival, we had a great time despite finding sand plugging up our noses, ears etc....Recovering from roadkill and livestock issues (I think I will have that ham sandwich now). females perspective of the Carcass Concert:

Firstly the trip there took about 2hrs.

The place was not well lit - being like farmland the ground was pretty uneven.

After being warned to wear pants and something warm (which I totally ignored) I fell in a hole (Please laugh - I mean really laugh) .

The loo's were about nearly a kilometre away - hence I saw more of the loo's than the concert - one great band (my opinion) ROSS (a black band).

Carcass did not disappoint - but they took until nearly 12:40 am in the morning to start playing. I was in the car shivering to death and gave up on trying to watch the band as it was raining and way too cold.

All in all it was for *Name* - so guess who will be paying for my support affected you (positively or negatively)!we can work around this problem. How is the search for *Name* going????

I went to the Alcatraz Reunion but no sign of *Name*.

Have you checked out the numbers in the telephone directory? Do you know what *Name*'s parents initials are??

Need them to search. Sorry to hear that you are not well, is there anything I can do to help find *Name* - please share your ideas...

I did not enjoy the Alcatraz Reunion as The Band -" Sheep on Drugs" was not very entertaining and all the people with exception of one guy were not recognizable (Emo's) you can check out pics on I remember Club Alcatraz site - Sheep on Drugs/Alcatraz Reunion - Hartiez. I enjoyed the evening out - and No I don't have any info regarding *Name* - but hope prevails - I think checking the Phonebook out anyway is a good start - *Name* is my Sister-in-Laws Uncle. *Name* spoke to *Name* last night and so far everything

seems to be on track for Saturday. *Name* will give you a call and confirm. Sounds great - See you soon.....! Sorry to hear about the loss of your Dad - our thoughts and prayers
are with you - Take Care I will be going to the Exploited Concert - I was not sure whether *Name* would be able to organise tickets - but he has - so I will look out for you there. What's wrong with your mom????
Send my Love, let her know I am praying for her. Hey *Name*, had a quiet weekend - I am glad to hear that you are rested - looking back I used to love watching my children as babies "the innocence" really touched my heart and I used to kiss them to bits, loved their perfect soft skin.
I had to laugh at your comment about 'holding out' and then getting 'sucked into FB' and that scary incident at that church - I will tell you that I have had many more but learnt one very important thing my faith is not in vain "Jesus is real" and awesome - He delivered me from smoking and drinking by His Grace and continues to love me through all my ups and downs.
The spiritual realm is not for the 'faint hearted' but I will always be Thankful to You for your support and willing spirit in helping me on my journey to freedom.
Bless You and Bless You again. I know how not having sleep can affect you, our son woke up 3 times a night every night for the first 16 months of his life. How is she otherwise? eating well? shots can also be tricky - normally they advise a bit of Panado syrup.
I am glad to hear that you have not changed - I remember you having tea when the rest of us where beer guzzlers - really impressed I was - I guess I wanted to be self-controlled on some level but just could not pull it off back then and you know that the spiritual problems I had put me on a path that I am so relieved is the right one - my faith and loyalty paid off in the end - Jesus is awesome - most people do not understand the spiritual - I had no choice - My life had to be cleaned out supernaturally as I loved smoking and drinking too much - The Lord in his wisdom had to rescue me from myself and evil. (refer to lights flickering conversation)
I also do not listen to any heavy metal or alternative music any more - once again - deliverance - I find that anything that becomes addictive is no good for the soul to be

honest a lot of the music I listened to made me very depressed or caused me to become agitated and jumpy - I am an avid fan of Tree63 local Christian Rock band (now in living in the states)
I hope for your sake there will be no smoking allowed - I also cannot handle smoke!
After saying all the above - please note that I still have my sense of humour and am learning day by day just to be me and not let others
control what I do and say - hard- but sometimes we all have to 'draw the line in the sand'
Sorry to hear that the shot was so traumatic - the bond between parents and children is in the heart - I am glad to hear that she is a good chucker -

---

my son gives me uphill *Name* says, not to worry she actually forgot about it.
She says there is no need for you to make it up to her. She's just grateful you thought of her that is what really counts. We normally don't have big parties for the kids anyway.
Are you enjoying being a Bookkeeper? As for my medications the well has run dry....I was given so many I could not keep up with it, treating ADHD is difficult and some of them turned me psycho...You know who got it don't you *Name* yes in the neck and everywhere else- Thank God we don't own very dagger like type knives' am discovering sides to myself that can be horrifying honestly brain chemistry is a delicate issue and Thank goodness my psychiatrist had the foresight to give me a whole lot of information and coaching lessons. I actually do know who I am Thanks to this experience and am learning to love myself as I am (not easy) but I Thank God for His Help in helping me discover that I am not a bad, awful person but am someone who needs help and understanding.
I don't know when I will see you but until then Take Care of yourself! Hey Lindsey,
Firstly, yes I have recovered although I do not know, what exactly I went to see a psychiatrist and he asked me who I am? I said I do not know? as a result I am on many schedule 5 medications,
As for *Name*'s her Birthday is on the 22nd of March and she would love to go for a facial with you (I will be so happy for her). As for me I will save until I can afford a full facial and then try to make a plan to go and in the mean while I will use Olive Oil to fill in the

cracks,
Thanks for your offer to twist *Name* Arm - But there is no need (I will spare you) I will make a PLAN, Lastly, dont worry about INFLUENCING me as you will have a hard time seen as though I am still trying to figure out WHO I AM Take Care of Yourself and Have a Fantastic Week Ahead, I will keep you in my Prayers Hey Same to you and yours. We have been doing home improvements and the place is a mess has been since before Christmas. I am so glad you found *Name* and once things have settled down - I will phone him. Enjoy your time with *Name* - Here's wishing you everything of the best for the new year Well, it took you long enough to answer. This has assured me that miracles really do occur. hope you and your family are doing well, until I hear from you again, take care and send my love to your family. Hope you are looking after yourself? I hope you have a great week ahead and that this year will be great for all of you....Your powers of observation are rather unique.
Of course the sixth cube had to go first. To sum up - A reduction in the amount of ice cubes ingested would perhaps be a temporary solution to Impulse control issues. It is the will of God that every believer has an intimate knowledge of Christ, against this backdrop, I will share what I have come to know through revelation. Our God is a God of relationship and Christianity is not about rules, regulations and condemnation it is about Relationship. Jesus is our Friend as well as our Lord and Saviour He looks more on the heart of the individual than on the outside appearance. Our relationship with Him will eventually cause us to become more like Him in every way, it is progressive. I personally consult The Lord before doing anything in this way I stay close to His will and Him. Keep in mind - Condemnation is from the Devil - Conviction is from the Lord and is always done with Love never leaving the individual with a feeling of despair or rejection. Read More I have personally experienced Deliverance from Smoking and was healed after spending 9 days in hospital with no result from the treatment that was given to me. I am in awe of the way The Lord puts people who are in a similar situation together to comfort and support each other- He is truly the Author of our Faith. He has moved on my behalf in many circumstances and reminds me constantly that His Love is UNCONDITIONAL !linked to my above reply - I want to point out that I have been in a church where I had many negative experiences eventually I left and decided not to join

a church until I had proper direction. I hand my days over to the Lord and Let Him do His Will...I through my life everyday - but I recognise that belonging to a Church is part of His will most definitely. Learning to get on with people in this way (church) is difficult and rewarding depending on what you are focusing on - your experience will lean to that end. I came across many angry aggressive women and could not handle them at first but The Lord worked in my heart (over a period of time) to love unconditionally and as a result I am more loving and patient. Remember God is interested in developing your character as I found out this does not happen when you shut yourself off from the church. People should not get in the way of our relationship with God if they do then we need to ask God to help us stay focused on Him! Not sure whether I should be concerned about...what was it again duck flu or frog flu or was it swine flu ..... dreaming about pigs not yet. Bacon, no thanks.....

Writer B
i forgot to ask you, but I owe you the 370 something which I still need to put into your account, but *Name* wanted to know if she could put R1000.00 in your account and could you give the AS$ to BG for *Name*'s present?
if it's a problem I'll find another way.
hey *Name*

damn just missed you online....
you landed already!
how's it like been home?
hey ya's

Name both of us are cool with either sunday, so which suits you best? (13,20) I personally prefer the 20, as I suspect a late nite on the 12th and early morning sounds daunting. heeheehee.
We can meet at my place and go from there.
I know this great spot that we can picnic at in the JHB botanical gardens.
you're welcome to invite anyone else.
Was thinking of asking *Name* and *Name* but I think they are still in ramadan and she may not be too charmed with seeing us eat while she's fasting....:) so will see.

okay speak soon. Okay confirmed - 20th at my place at 11:30! And we'll go together.
i have also asked my sister and bolien you met her the last time we were meant to go for a picnic...
Alas this will be the real thing.
see ya soon!!!
B yip! We're still on!
But it seems it just the 3 of us, *Name* cancelled on me last minute and *Name* has Eid.
so i'm gonna get bread and cheese and some parma ham, with a tomatoe salad.
Do you and *Name* wanna bring drinks and dessert or fruit
and I think we're kinda set. Unless you want something else for the picnic?

B
hey you

you been a stranger... even on facebook!
how you been? what you been up too?
did you get the invite to another WPoint party?
you have to join, it's this weekend, and maybe we can catch up?

B hey honey bunny!!!

A little birdie told me so sad news!!! is it true you're leaving us????
y y y y y y y y !!!
we need a catch up?

you still living in fourways? maybe a drink after work?

B sniff sniff sniff...:(

Okay so when you available?
Otherwise speak to T this weekend and see what date is good for her to do dinner next week??? wow you actually got onto facebook!!!
i'm impressed! Well now you can share in the goss...:)
B

hey honey

if you have an intermediate on Monday, are there enough men around????
would love to join, but will confirm this weekend with you.

B
Thanks for the insy winsy tini tiny bitsy witsy piece of info....
heeheehee. What's the good in knowing a Mexican if he cannot help...
heeheehee...

So what is with your messages and you bitchin all the time? you know it takes to TWO to communicate... you can drop me a line or two too? you don't have to wait for me to mail you to say hi! But being the bigger person I shall clean the slate and start...

Alas how have you been? what you been up too... mischief?
You still working from Mckinsey? Are still crashin with a friend?
Done anything exciting? How's your spiritual quest going or is it just a vocation change or a distant thought?

Things here have been a bit crazy, with my company's FYend work has kept me and still keeping me on my toes but socially the group has been busy changin dynamics, so it's been a laugh...

Been kinda weird lately though, in a different headspace... guess I know I've gotta make some changes and choices and in times like this the mind is always like (excuse the analogy) a toilet bowl of swirling shit, that desperately needs a flush. but such is life... so hopefully some new developments will take place...

Other than that Tango has been my other love and I'm loving it, found a new partner the night of your farewell, and so far so good. We've been progressing really well and I'm really beginning to enjoy it. It started of a bit shakey, as I was dumped very abruptly and then had to find a new partner and start again, which was kinda disruptive and frustrating. But I think once you get past the initial hurdle Tango is for life.

Tonight I'm off to a Tango Melonge (social) and thereafter friends are meeting at the new renovated Katzy's for live jazz, so should be a good night. We were there on

Thursday last week, and the band was awesome played all the 70 rnb stuff. Really good!

Okay so now that should stop the bitchin, and hopefully get a DECENT response from you... So don't be cactus prick and blossom yourself to a cactus flower...

Have a great evening/day whatever applies...

B
I see you're already on holiday! i'm sooooooooo jealous!!!
Are you island hoping in Greece? or stayin on the mainland?
Well now you should have plenty to tell as life should be adventurous at the mo...

I'm sick at home and trying to get better, which is far from exciting so tell us some stories...

Where you off to next?

Enjoy!
ok i'll wait for this 'more'...
tap tap tap tap...

Yeah better a bit, but glad for the time to chill. was in bed a lot and caught up on Dexter - 2 seasons - fabulous!!! Glad I didn't go away with the rest of them.

so happy travels...

why are you going back to NY and then travelling back? no maketh a sense much???
Hi aunty *Name*

Sorry for the delay. Please check you mail as I have sent you an email with an attachment. I wasn't quite sure what email to use so I mailed to the mail on your facebook and another gmail account.
If you do not get it please fb me and I'll resend the mail on fb without the attachment.

Regards
Name Hey *Name* boy

how are you??? what you been up to?
besides raiding ol pics! you were such a little cutey when you were a kid...
...what happened.... heeheehee.

We were meant to go and have sundowners at the westcliff this sunday and *Name* kept reminiscing about you your stint with Clint Westcliffwood...:) She's missing being

naughty with you...:)

tell us some tales...

B
Hey ya

Happy Birthday!!!
Hope you're having a fabulous day, I know you gonna have an extended birthday celebration with a lovely visit from the one you love...

May the time be cherished and special.
May your birthday wish come true!

Love
B
Yeah you are blessed. Added to that you have he weather on your side.
I suppose being in Canada has another advantage - you get to have your birthday in the Summer!

So how exciting must things be for you right now? What surprise did you have up your sleeve for *Name*??? we're all curious!
Meanwhile I hear that you have lost like heaps of weight too! I'm so jealous that I'm seriously looking at this diet you guys are on...

Ok, chat later.
B

don't know about living the dream, however am havin fun...;)
how have you been? what you been up too?
yeah some dodgy pictures, need to be careful. heeheeheee

have a fab weekend.
Hey good on ya

glad to hear that your crutches have left you, lets hope it also makes for a fabulous day!!!
Okay, again so how do you know Bg?

The St john's boy look, arrogant bastard....
heeheehee.
only kidding, just that I was at your sister school...

what you been up too?
hey ya

Yeah, no probs, sorry I didn't reply to your last mail.
Fri and Sat were rather hectic days and didn't allow me to get to the internet. Hope you had a good Saturday nonetheless and that it was had nothing to do with your domestic issues...

The band was good, we kinda got there rather late and only got to hear like 3 songs. We kinda had an incident with the police coming down a oneway in town... Alas got there safe and sound and without a fine, so all's good.

B
Hi ya

Didn't see your directions till monday. So used a friends gps that went a stray... and end up down a one way in town...

Saturday a bunch of us went and spent the morning painting a primary school in Soweto, which was great fun, loads of colour and good for the soul! After which my friend *Name* and I dashed to Muldersdrift to meet another friend of mine to see an outdoor exhibition. And WOW is all I can say. It was on this amazing estate that apparently is an artist retreat. The exhibition itself is of really famous SA artists. Additionally we had a guided tour, by the curator who was fantastic with his anecdotes. Then to end the evening we went to see a Spanish band in Brixton. House of Nsako - Cute little venue and the music was not bad.

So as you can see it was a good hectic.

This week is kinda hectic too, maybe next weekend, oh wait away that weekend... i'll keep you updated... heeheehe
Man you crack me up!!!

Funny that most of the people responding are family, so they're all going to be receiving the comments, and they are going to reel with your comments. So I'm anticipating a string of queries. It's going to be sooooooooo funny. I'll keep you update with the funny ones!!!
You have no idea how I'm canning myself at the mo...
Thank you for making my day! It's so what I needed!
the Universe is good to me and is hearing me all the time!

So other wise how are you? long time no hear.
I must tell you, I went to Moz for a long weekend and fell in love with a local artist there. Have you heard of *Name* ...Oh no I forgot his name...
Will have to get back to you on that too...
But he went to the coconut club and he was performing and man I was swept of my feet, I guess only because it was in Portuguese. My friend said that the words are absolutely ridiculus, but I didn't care...
heeheeheehe.

Thought you'd like that piece of tit bit... heeheehe

hope you have a fab day tomorrow!!!!

B
mmmmmmmmmmmmm! is he cute????
heeheheehee.

Apparently Catzy's has re-opened and been refurbished so Al's grill house is a big winner - Great Steak and live music and entertainment.
You have to book for Al's grillhouse as they are always full. And you can walk over to catzy's for dinner after.

Is this person young or old? What kinda food?
There's a new Pigalle at Melrose arch?

Otherwise you mentioned Metro, it's very black, but it's' cute. I'd suggest though to go to O'gallito's for dinner and then go next door to Metro for drinks.

Also when are you planning on going? Sorry not o fay with the rugby details? Where you go for drinks on a particular day makes a difference.

Meanwhile how are you? When are you having your house warming party??? I'm still waiting for the invite.

B
hey honey

What exams are you writing???
That's cool, I'll see you on Sunday the 21st regardless.
Otherwise we'll never meet up.
Where abouts do you stay, do you have any suggestions or preferences? And do you wanna do brunch or sundowner drinks?
hee hee hee that is funny. hope your son's fine now. Just stay well Mommy, can't be letting you get sick too.

I'm kinda easy. I'm the Melrose area. Maybe we can chat about this later, as I was thinking of somewhere outdoors in the sun. But then with the current weather we are experiencing, i'm thinking a fireplace would be great... so lets see what the weather is like closer to the time.

Writer C

You're looking fab in that pic girl! Have an awesome time this w.end :) and i want to hear all about it next week x sweet dreams

Hello (stranger??) Is this r molin that i used to work with? Apologies if not-there was no profile picture so I wasn't sure to reply.. Thanks

Hey *Name*, all is going well thanks :) Though i almost died today! Had a blow out on the highway. 2 guys (i'd call them my angels) stopped to help me - done in 10 minutes then i was back on the road. Thank the Lord ! very scary stuff! For the record - i can change a tyre - but its gr8 to not have to :)

How you doing? Jetting about by the looks of it - Zambia and Syria, sounds exciting. Hows your brother doing? I chat with *Name* now and again - but haven't seen anyone for a zillion years. Facebooks pretty cool for hooking up with everyone.

Right - im outta here. Hav a better eve

hey there Stud ☺

Firstly i dont know why i am up so early! ! :) like you, i was also drinking wine last night, and seriously i'd much rather still be sleeping! :) i made a nice thai dish for a friend-but used all the veggies she doesnt eat! So that was hilarious! Though i felt so bad!! Should 've made lasagne- everyone loves lasagne :)

I'm so sorry about the heart break! Breaking up sucks big time. I would never even wish it on a worst enemy even. You guys must have been very close, together for a long time?

You've got the right idea- keep yourself occupied, meet up with all your mates and wine! Wine! And more wine!! :) a bit of gym also helps i find-but not too much of that! :)

Cypres sounds soo cool! Hopefully you get in alot of beach time! Do you get time to do a bit of sight seeing when you go away on these trips? Hey in syria you can find a lady to come back with ya to do all your cooking! Just take some camels for bargaining :)

Yeah im also tiring of my work-i sell insurance-been doing it for 7 years- i love it, but of later i get a bit bored. In fact i recently chatted to a colleague about her job which is in training. But not making any big decisions yet-still have some things to achieve at work.

Shew , im sure im typing an essay now.
So to finish,pleased you're okay after your accident, and yeah-its gr8 that no one was hurt. Motor bikers are nuts taking such a chance-what protection do they have against cars out there.

K, see ya and have a gr8 w end!

Name honey - you're looking so hot you are! single life is working well for you for sure ! any man that sees ya - know you mean business :)

Wow hon - credit control manager - GE was a good place to start :) wow - s you're good looking and wield alot of power :) Professional too - not mixing business with pleasure :) you must have learnt that from me *haha -remember *Name**

News - well work is going very well. Though i worry about the commission changes going forward.
I bought a 2nd property - which i plan to move into in the new year . Then you will have to visit!
Fam is well and amazing as always. My folks are away at the coast for a few days - they need a break no doubt.

Newbreaking news : I' am trying to cook more often!! so working on a great Thai dish - the type that 'll make a man wanna wisp me off into the bedroom or fall in love with me- mabye not in that order :)

k hon - mind yaself!
miss ya
x

hello you beaut! How you holding out there? Did you enjoy your party w end at all? How are things with f? And d? Just thinking of ya and hoping you re feeling better x

very much alive! Just had sushi and creme brulee :) yum yum yum! Did you watch district 9? Effects good? I guess you have no news then...

you kidding! Bad Timing :) off to watch it now with my cousins. Later

so .. Where are things now? I think its some cheek he has told everyone about you sms d.
I knew this would make head lines. Not to keep score, but what about his lying and bad habits-unfortunately not the kind of info he would divulge. Ag, one way or another i hope you guys sort things out. Please shout if you need to chat. Though i do realise this is for you guys to resolve. X

You should maybe leave him for now..? F likes to play the victim. You think he'd realise he needs to get his act together. Far easier to carry on like he is the one that's been done in. Hon, leave d, it gives f more amo. Im good. Just had sushi with *Name Name Name* and *Name*, then funny movie ☺

was coming from my place to be when i passed ya! Cant wait *Name*! When i do get round to moving in you guys must come round for home made biscuits and coffe :) Hope all good your side. hey dont you have a wall on fb?

im chuckling at your comment about having a facebook consultant :) wine sounds fab! See ya then! It was only matter of time till I heard Boney M carols flooding the mall corridors

hey hey handsome! So pleased your exams over! Congrats! How did it go? Mine are going shockingly- why cant we just rely on our good looks to get us by? How you ? How s ya mom? X

---

hi *Name* sorry about that cut message.. Im on face book between studying chapters. Time is precious as you know. So we all well here thanks! April was a total pig out, with easter and other birthdays. The kids are growing, i think i am too,and not just side ways :) pleased c t is working out for you. So pleased to hear it! Plus its fab reading inspirational status up dates! How are you adorable children. Send them a huge squeeze from me. Take care *name*. X x

hello hello :) so pleased you got in touch :) talking about seeing asses- have you heard about the change in commission legislation...? Its halved on investments! So yeah tough times ahead in insurance :(
As for contacts at s a b, the last guy in any sort of influence left last month and is so difficult to get hold of. Is there anything in particular? A friend of mine works for the bank and is looking for project management candidates.. You can send me your c v. As for my uncle please would you ring him on 0726834009. I must be honest i was quite useless in running with the details. But please call him. In todays market he d be grateful for the odd job on the side. Right i.ve had a huge pizza and two glasses of wine. Must go to sleep to get some sleep and wake up early for gym. Guilt has got the better of me. Take care and all the best with settling in :)

cool dude. Send me his details. Could go tom or on w end. Shall i drop it off at simone after wards?

thanks *Name*. Been thinking i should change it.. But maybe not! You not married yet ;) hope all good

Hey Hey

So did you get my other message....? let me re-type it. Hmm.. took a sincere break over december - wow, great to have all that time to myself doing as i please! back to work and i've been off my feet. Studies have started so weeks filled with study groups. Just a quick hi back! how things with you

Heyyyyy!! i've been soooo crap with FB. admittedly have loved not being behind the computer - my main frame burnt out - then it was holidays, so i had every reason to spend less time on the internet and the holidays were GREAT!

Congrats on the quickest time. 4th place is good dude! - this is an international race right.. I was at Monte Casino some weeks ago and my younger cousin showed me the indoor skate park they have - bit weird.

bummer about the crash! anything broken? I hate injuries - really wastes an opportunity. you must have been pretty pissed off.

I work for Liberty as a consultant - yeah - being doing it for 6 years. I was working in HR - but it became so political moved into finance. My degree is in Finance and i study quite a bit for work still. i know - BORING!!!

haven't been riding for long - couple of years, may go to northern farms (near Magaliesburg) on sunday to do an off road track – about 32km.

saw *Name* about 3 years ago...?? she studied psychology, left to do hairdressing. She is looking great! she may be married at this stage.... ??
we didn't stay in touch after meeting up 3 years ago. you know how it is... passing by ;)

How were you holidays? things going well? Anyway, going to take advantage of the electricity whilst it's running - must do some work... :(
have a good w.end!

I think it's hilarious that you think i'm abroad and I think it even more hilarious that you think *Name* deleted us as friends.

I did notice i was down 1 friend, but it wouldn't be the first i've pissd someone off and i doubt it's going to be the last. gosh, i think the day you and i stop annoying people, would mean the end of our existence :)

I think *Name* was doing some changes which resulted in us being "unfriended". For some reason i recall him changing a name, I suppose he changed it back and in so doing - it appears his unbefriended us. Who knows? I'm pretty sure he still likes us as his mates. I sincerely doubt he was offended by that :)

Sure - so you need my emails - the black and white of my "blank and swipe" (mind) :) I'll arrange that for you. Just please bear with me whilst i arrange that for you :)

Hope the fam is well! It seems you're loving being a dad
Take care
Name
Hi I'm so sorry to hear about your dad.
how is your mom holding out?

Sounds like you've had quite a bit going, flu, sleep deprivation etc..

be lovely to meet the fam :)
will chat soon
Hey There

Yip my fam all well thanks.
Next sunday is pretty tought for me - i'm doing a cycle - which will probably leave me **** for the rest of the day :)

I did go out with him for a few years :)

Cycling this sunday only. Will get back to you on the pain :)

I see you're starting defence classes - that's awesome. another friend of mine is doing the very same thing.

Will be in touch then
Cioa

Really! i'm a numbnuts! I shall send you more of my toilet reading :)

I tried ringing you at that 615 number - it just rang out..

sh!t you're kidding! sooo sorry!!!

willl send it pronto!!

So did you get my other message....? let me re-type it. Hmm.. took a sincere break over december - wow, great to have all that time to myself doing as i please! back to work and i've been off my feet. Studies have started so weeks filled with study groups. Just a quick hi back! how things with you?

Writer D
Name!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! Hey how are you!? I believe congrats are in order... when are you due?? Or have you have baby already!?
when dude?? need to order a day or 2 in advance... when you up here??
when dude?? need to order a day or 2 in advance... when you up here??
hey *Name*!!! wow... it's been a while!! how are you?? How is *Name*!? put up some pics soon!!!
why hello madam!! it's been a while... i think at *Name* place... when she was that ass!! :(

How are you!? Where are you!?
hey *Name*.. how are you? back to the grindstone back in ct i take it?? hope you are taking care... *Name*
hey there!! wow.. long long time! what are you up to?? and where are you? still in SA?
hey there *Name*!! How are you!? what you been doing..haven't bumped into you in ages!!! did you enjoy sex and the city last night at monte!? :) bumped into *Name*... but didn't see you.. think we had gone to get something to eat!!!
huh?? where on earth did you get that idea from??? nope not engaged.. sorry to disappoint!!! so is it that bad being back??? your hubby looks like quite a cutie!!!!!
hey *Name*!! how are you!? It's been a while! ;)

Where are you now? Married? Kids?

I'm still in jhb, living in the Sunninghill area...not sure if you know the area? Living in sin with my boyfriend of about 4 years...not married...no kids... :)

Working as a Business Analyst in the IT field for the past 8 years of soo.. I know.. boring... well, it pays the bills, and the company that I'm with now has flown me to London twice this year... so I get a couple of free travels!

Tell what you've been up to these past... 15 years or soooo!!! Nice to have found you on FB!
hey there... thanks for the chat last night.. I don't think I've yakked soo much in a long long time!! :D

Looking forward to Uvongo.... let me know about Sunday dinner? And then just bring me the keys... and what ever you need to be taken down to Uvongo.. just not the kitchen sink!!! Haha!!!....

And I'm sure there's something else I wanted to remind you of... but I've forgotten... :P
Name Name Name... how are you!? Moved in yet!?? :)

Well... if you're still looking for a washing machine, let me know. The one I am selling is

an LG frontloader, 7,2 Kg, white... R2000. So let me know!!!

Otherwise.. what you been up to!? Have you seen cel yet?
She's a beauty Mi.. congrats!!! How are you doing... enjoying the 3 hourly feeds!? :)
hey there... how are you doing? Well, my next trip to the UK is up already!! Unfortunately for a short time though, and specifically for training! :P Anyway... I arrive on sunday, and leave on friday!!!

Just in case you are available, and if I can get away, it would be great to catchup!!!
hey how you doing??

Will be in london on sunday!!

i take it you're still sleep depraved!? :) Said to *Name* that the 2 of you should move in then at least you'll have each other to keep company in the early hours of the morning!! :)

Anyway.. chat to you when i get back!!!! post more pics of *Name*!!!
Congrats on your masters.. you must be sooo relieved!?

I'm back in SA now.. only went for a week's training...and a little bit of shopping! Okay.. a lot... so much in fact.. i was overweight on the way back... :(

How the hell are you!?? Where are you?? What are you doing?? Babies...?? :)
CONGRATS!!! wow... cool... Nelius must be extremely excited as well?? so.. and october/november baby?

I'm at a company called Softscape, a US based company, who develop HR solutions. also been keeping me out of trouble and quite interested! :) Been with them for about a year now.

how long have you been at KPMG for? You left BB&D and went to.. and health care place ja??

Is *Name* still at BB&D?? I believe there's been quite a few...'happenings' there since we left!! ;)

otherwise, can't complain.. been living in sin with *Name* for about 3 years now... we bought a place in the Sunninghill area about 2 years ago...wow time flies! And you? Still in Centurion way??

Good to be in contact with you!!!
Wow... how long has it been since we last saw each other!? Okay, so I left RMB, and joined Discovery...and they didn't really keep me that interested...so as you said... I moved on! :) I moved onto a company called Softscape. They're a US based company, who develop HR solutions. So.. something quite different for me! But... it's been just

over a year.. and they're keeping me interested! ;)

Haha... ja, I started going out with *Name* basically just after I left BB&D! In fact, i took him to the year end function at BB&D!!!! So, ja, been together since then... he moved in with me, as I was still staying in my one bedroom apartment in Sandton.. then about 3 years ago, we decided to look for a place... and so 2 years ago we moved into Sunninghill!!! Not a huge place, but big enough for the 2 of us... 2 bedrooms, a study.. and open plan kitchen, dining room and tv area! A tiny garden.. and a double garage!

I just came back from the UK...went on training! Bought a LOT of stuff... was overweight!! But I caught up with *Name*.. do you remember him? Not sure if you remember his fiance/wife/exwife now.. *Name*!!? Yes.. ex wife.. can you believe it! Anyway.. soon after they got married they emigrated to the UK... and basically ... i think it took it's toll on them.. and they split.. I was VERY VERY surprised. :(Oh well.. but *Name* seems to be doing well..

---

and enjoying bachelor hood.. in a good way though!!! He was always a sweetie!

Let's see what else.. Oz hey... I think you... and me.. and about 80% of other Saffa's that want to leave the country... are thinking about Oz!!! What are your plans for that?? Soon.. obviously after baby is born?? How is baby doing??

You know what.. we should meet up sometime for coffee!!??

Anyway... send my love to Neels... if he EVEN remembers me! :)

Chat soon!
Hey *Name*!!

Yeah.. .life seems to have caught up with me as well!! In London at the moment for training!!!! Will see if I can organise something when i get back... next week!!!

Hope you sell you house soon! It's difficult now with the interest rates...etc etc etc... more of a renters market now!

Later *Name*
Hey Aunty *Name*... welcome to facebook!!!!!
hey.. i'm in london... till friday!!!!
hey there... i'll be there on sunday!! fetching me from the airport?? plane lands at 7am!! :) Kidding... got some other sucker to come fetch me!!

Well... let me know how your sunday evening... or any evening during the week looks?? i'll have my SA cell on me...so can text me on that... *cellular number*. I'm assuming the number you have on FB is ...your cell number

later
cool man... well I'm staying in Camden, at the Holiday Inn Camden lock... but i'm pretty savvy about getting around with the tube and all... but yes... chat you on sunday!!! not sure what room i'll be in, but here's the hotel number.. it's probably cheaper to phone there and get through to my room!! *Telephone number*

looking forward to catching up!!
and there i though you only had to feed the little thing every 3 hours during the day...only!!!???

She's gorgeous!!!!! So 'perfect'..little round face...smiley eyes..

Perhaps you and *Name* should move in together and both get up every 3 hours for feeds!!! hehehe!!!
hahaha... shame man.. i feel for you and *Name*... she's just asked why sleep can't be 'accumulated'.... so i'm guessing she's getting 5 minutes here... 10 minutes there... another 5 later... post more pics!
Hey *Name*,

good to have you on FB!! You STILL look the same!! Never age!!!
Hey you!! How are you doing!!! So what are the chances of you coming over with your mom and sister at the end of the year!? Or do you have school!?

Later
Name
Wow 8 years!! Long time! To think, i'll only be getting married in Dec... first marriage.. no kids.. ;)

Do you remember *Name+surname*? She is now *Name+surname*... she's married to my fiance's cousin... small world... or just the small chinese community! I also stay in touch with *Name+surname*.. she's now *Name+surname*.. she has a set of twin boys.. about 4 years old. *Name* has a boy (4), and a girl (2)...

What business do you run from home? I would LOVE to work from home! Really itching to do 'something'.. not sure what!!!
oh my goodness... i wasn't sure at first.. but wow.. you look stunning in your profile pic!! you must put up more pictures of yourself.. and your little one I'm assuming! I think I'm just picturing you with the curly curly hair back in the day!! haha!!

So, I'm assuiming... married with children!? When where.. who??

I'm engaged, planning a wedding for the end of this year! Marrying a local chinese south african boy... *Name*... 2 years older than me... hopefully a good match!

Otherwise.. ja.. left McAuley.. went to another school, finished off there, went to Wits,

got a degree in Economics, Insurance and Computer Science, and have been working in the IT world for the last 9 years or so as a Business Analyst/Project Manager... Phew.. and still living in South Africa!

It's sooo nice to have people on facebook that i haven't seen or heard from in ages.. like you! wow... let me know what you've been up to! ;)

later
Name
Hey *Name*!!! Hope all has settled back to 'normal' there... if there is such a thing? Thanks again for having us!! It seems sooo long ago now...

Could you help tag some of the people we met at your party? Especially all the people that know our parents... ;)

Anyway.. will chat soon again!
Name
well.. we started in Toronto.. then we moved to San Fran... then LA.. then drove up to Las Vegas.. then flew onto Vancouver.. then ended in London.. then back home.. phew! :)

Yes... will definitely make plans when ... either we next visit.. or when you're back here!!! Looked like you had a ball of a time.. nobody has changed!! :)
Name+surname!! Oh my goodness.. how are you!? Where are you!? Living in Melbourne I see... according to Facebook that is! What are you doing there!?

Good have contact on FB with you! Hope you are keeping well... let me know what you are up to!!?

Name
yes yes..i know you are a 'god'... but i hope all is goood!? :)

You too hey... how often you change jobs? I think this one is my longest.. and I'm starting to get itchy feet... been 2 and almost a half years!

Otherwise, doing well.. thought you would've emigrated by now?

when when... lunch sometime? on a weekend... i tend to actually work during the week.. i know,... not like me.. hence a change in job scene is required! ;)
Hey you...

We'll be there soon!!!!!!!!!!!!!!!!!!!! You going to be in town? :)

Anyway, would be nice to chat anyway.. so my friend's home number is 44 208 989 5963... will be cheaper to call me at my friend's place, rather than on my cell!! ;)

Anyway.. speak to you sooon!!!

Name
Hey there!!!

I'm popping into town... arrive on Tuesday the 23rd.. and leave on Friday the 26th.. yes, it's a short stay, only 4 days... or 3 and a half days! We're staying at a friend's place, they live in South Woodford.

Writer E

Cuz I'm heading off to bed.
Will rattle thw windows tomorrow with EBE 1947!
Take it easy
send my love to your mom and dad
Big hugs
Name

P.S., who is marketing your songs? besides the DJs at the goth clubs? Have you consdiered contacting Barney Simon at 5 FM (not sure whether he's still at 5 FM - but he was their DJ and played all the alternative stuff and you could send him demos and he's play 'em

Hi
I don't know where the time goes... one minute I'm getting up for work and the next it's time for bed... whew!!! I saw some piccies that your biomom put up - they are lovely!!! I see you graduated... What did you graduate in?
I spoke to my mom (Aunty *Name*) the other day - she sends you her love... At the moment she is battling to get on the Internet - but said as soon as she does she will try and get on facebook and tell you lots of stories....
Take it easy!

Glad to hear you've left those poor plants alone (hee hee) ummmm not studying this year... but may do next year.....I'm doing very well thanks.... Well I was in England (for nearly 8 years whew) and now in Aus. (nearly 2 years here whew)..... ummmm am I loving it? It's hard to answer that one.... as there are parts that I enjoy i.e. being with my boyfriend, *Name* and our cat *Name*. The people here seem to be friendly and quite up beat.... But, on the downside I miss my family and friends in SA and UK tremendously... And you? What are you up to now a days?

Aha.... yip makes sense why you would go with that... I understand totally... the thing of you are where you are meant to be....
Me? WEll I met my boyfriend/fiance in the UK. He's Australian and was coming back here and asked me if I'd like to join him... I said YES!
What are your little girls names? I was devastated when *Name* told me about your oldest daughter having cancer - but she's a strong little fighter!!! and am relieved to hear that she is now ok. They definately would get a lot more outdoor time here....and so would *Name* (It is *Name* isn't it? - if not my apologies). Have you had a look on any of the job websites? there's one I use www.seek.com.au also www.careerone.com.au.
What work do you guys do?
Also just in case you're looking, there's a realestate website called www.realestate.com.au.... You definately get a lot more for your money here compared to England. Whereabout are you in the UK? Have you gone back to SA at all? And, if so what are your thoughts?
Have a great weekend

Hello
I'm doing very well thanks? and how are you doing? I'm still in Melbourne enjoying the free life...can't believe I've nearly been here for 2 years.....Are you still in the UK? Fingers crossed it all goes through... Why have you decided to leave the UK? and why Melbourne?
Cheers

Sorry cooks didn't mean to make them sound bad.... but just got quite a shock... As long as the two of you are looking after eachother (women love breakfast in bed Sunday mornings - hee hee)
The bride test is something my younger niece sent me - as she'd taken it she's only 12 so I don't think age is a factor.... in other words I have no idea....
Name is a mathematician - he does stuff with numbers in a pure sense ... he works at the uni here doing high-tech research and lectures...
Right where to begin with the famdamily....
My mom and dad finally got divorced (after 40 years) - they have now gone there kind of separate ways... Both now live in England....
My sister *Name* is married to an Englishman - she also lives in England - No kids or pets.....
My sister *Name* got divorced - he woke up one morning and said he couldn't go on like this and moved out that same day - Fortunately she is strong... and has moved on .... my younger niece is still battling with it)... So that's the family...I do try and stay in fairly regular contact with them... Although sometimes they can be vol kak and then I don't.
Glad to hear that you have finally made the move and started your own bussiness!!!! Hooorrrrraaaaayyyyy!! And what's this with the race bikes - you mean motor bikes? or bicycles??? Tell me more news!!!
Anyway dinner is ready so I'd better get going,
Big hugs to you both

gee, sounds quite hectic! mmm a monk eh? Maybe a monk-ey (hee hee)
Well I left SA about 9 years ago - lived in the UK for 8 years... Studied some more... got my PhD in Psychology (Yip it's official I'm now a Dr. *Name* no longer a Miss)... Then I met *Name* in the UK (he's Australian - asked me if I'd like to join him in Oz.... and now I'm now living in Australia with *Name* and our cat (*Name*) - it's fantastic!!
Have you been in contact with your family at all?

How long is he going for? Why didn't he invite you along too? Nice to catch up with your other flowers... i mean buds... Is *Name* and you still in the same class? Do you and *Name* ever get in touch?
How you feeling about A, etc moving out? *Name* seems to be stressing about it...I think it'll be much better for the 3 of you - what do ya think?
Cool about the band - have you guys got any gigs lined up yet? I remember there used to like a competition on called 'battle of the bands' in SA... that might be something worth investigating? So what songs do you do? Are they covers? or originals or a mix?

Why was the *Name* ignoring you? Tell her Granny *Name* says she must give you a big wet kiss!!!!
Oh I've got a BIG black fur ball looking at me going 'mmmmooooooooooommmmmm' I guess I'd better go and feed him otherwise he might waste awayyyyyyyyyyyyyyyyyyyyyy
Love ya lost

How was Gold Reef? Did you go on those scary rides?? Did you go with your mom? or dad? How's things going at home?
Me? I just stayed in bed with Homer as I've had a baaaaaaaaddddd cold achooooo (excuse me)... I did manage to do a bit of gardening today with Homey's help of course.... I dig and plant... he pats down the soil... Seriously *Name* he does... the cat's a genius!!!hee hee
Anyway cooks *Name* just called me for dinner....
Love you lots

A. *Name*
P.S. I heard from nana the other day she sends you lots of love and hugs
Whew! Gee whiskers *Name* just reading it makes me go whew!!!
Sounds very busy!!!! where you going to have it?I'm doing very well thanks... I'm now living in Australia with my partner *Name* and our cat *Name*.... Glad to hear that you are both doing well! Many congrats on getting married too (I did send a message via *Name* to congratulate you - not sure whether you ever got it???) mmm I can see *Name* still working with people - and I think he would make an exceptional nurse! And you? what you up to?
Big hugs

industrial disease??? like bosses and things eh?
Yes Australia isn't just down the road.. (is that that rood?) - in Melbourne - we've been very fortunate to have been affected by the fired directly - but there are many places where we have previously visited that have - it's really dreadful.
What's *Name* up to these days?
Hugs

Glad to hear the parentals are good. Please send them my love and a BIG hug!
Is there much money to be made in books in SA? or music for that matter? for me it was never really a country that encouraged the arts.
I noticed that you are a fan of Tesla - among others - have you read the book the Prestige?
As for *Name* - what do you mean we don't want to see her with her man? Is this *Name*? or another? and what on earth is this emo thing?
Down under is pretty awesome - minus the bushfires - we have this thing called freedom - have you heard of it?
Lots of love

gee in my day they were just called teenagers - Now you have names for different types of them - *Name* does smile - actually she has a fantastic smile... As for issues yeah I think the biggest one for her was the day her father buggered off with another woman..... But, I believe in Karma - and one day....The Prestige is a fictional book - it's also a movie... another good movie you might enjoy is the Illusionist.....
FUBAR???? Haven't got a clue....
Big hugs Cuz

mmmm maybe it's a bit like ice-cream or chocolate hey? we know we love it but we just don't know why...But for those people who've never tried ice-cream or chocolate - trying to explain to them why it tastes SOOOOO good... can be rather difficult....
I guess we just confused about it all *Name* - please bear wif us ag man (hee hee)... I thought maybe you weren't happy at home.... Are you happy at home? or did you prefer it when there was just mom, *Name* and you there? (and *Name* and *Name Name* of course)...
Moving onto other things - you know in my last e-mail I mentioned An Interview with a Vampire - well I noticed you signed up to Queen of the Damned - well that's in the same series of books - I think there's the trilogy - Interview with a vampire; The vampire lestat and the queen of the damned...
Anyway snoeks tyoe some more soon and let me know all your news
I love you

ummmmmm that you love me???? that I am your favouritist aunty in the whole wide world???
Just know that if ever you need to chat, I'm always here for you ok?

How's the dazzling *Name*?

Yip I am indeed working in my field - no *Name* not my corn field....
I did hear the pitter patter of tiny feet the other night - but when *Name* and I woke up it was a mom possum with a baby possum on her back in our bedroom.....it was magical!!!!
Weddings not yet.....
Well done *Name*!
Have you guys tried again for a baby woekie?
Have a great day!

what things happen in corn fields then? besides crop circles for UFOs?
How's the writing coming along?
ahhh pizza - still looking for a decent place that sells them here ... mmmm
So did *Name* have her baby?
ha ha I can see *Name's* face now.... and hear her too... ha ha ha
Shame did *Name* ever have counselling for her miscarriage? and you too for that matter? I guess it will just take time - and when she's ready and when you are then I'm

sure it will be the right time....
I worked for Melbourne university as a research psychologist - I was involved in a clinical trial assessing the effectiveness of estrogen in women with schizophrenia - Very interesting research - Problem was it was too far from home - I had to stay down in Geelong during the week and then come home weekends (ok usually I snuck off early on a FRiday shhhh don't tell anyone) - not conducive to a full time relationship... so they wanted to renew my contract I said no thanks...then mom was here so decided to have time off to spend with her and now starting the process of applying to universities and research institutes and hospitals....
Are you still in Newark?

Writer F
I was in Dubrovnik, but also went to Bosnia for two days!!! Dubrovnik is the most stunning city I've ever seen, absolutely awesome. So much to do - from exploring the old town, to island hopping and soaking up the sun on the Mediterranean beaches. I highly recommend it :)

I went to Lapad and Mljet - all the islands you do as day trips - not far at all. There really is so much to do, and there are also beaches in Dubrovnik itself. The beaches are pebble though, only some of the islands have sandy beaches. Some also have wonderful historical sites - churches, buildings - and some just wonderful natural phenomena. Google Dubrovnik and the islands and have a look. I cannot recommend it enough. Its also not expensive - prices comparable to here if you convert. Here is a link to a travel piece I wrote for one of our publications - it'll give you an idea

Hey *Name*

I am well thanks; I hope you are as well :)

Sounds awesome, count me in. You can get me on *e-mail* or *Cellular number*.

I don't really keep in touch with anyone besides those we both have on FB. Loraine moved to Italy and I don't know how to get hold of her anymore :)

Hey darling
I will discuss with you at work tomorrow xx

:) I'm glad to hear it.

As long as you are happy, that is all that matters. I imagine *Name* must have been a handful to say the least.

Nice. I love CT.

Aaah I'm well, plodding along. Not much changes as you know :)

Give me a shout if you come up to Joburg - I'd love to see you!!

:-) that's no good, you need to have fun to make it all worthwhile.

I sympathise with your mom. I also have a lot of neck problems, and I know how painful it can be. I'm glad she is feeling better.

Chat soon, and take care

LOL Dubrovnik rocks, very bleak to be on my way home now :)

Pyramids are good, not seen them yet, but have them on my list of things to do :)

Hey luv

I'm good how are you? How's the pregnancy going?

Dubrovnik is by far the most beautiful place I've ever seen, I absolutely loved it and would highly recommend it. It’s really awesome.

I didn't book it myself as I was there on business, but finding places to stay is very easy. There are tons of them :-) just look online. If you do decide to go, find somewhere close to the old town.

xxxxx

odd that’s the number I used. Mine is *cellular number*, can only receive sms, no calls. Here till sat, if you want to get together give me a shout.

:-) I am so jealous, I wish I was still travelling, getting to see all those wonderful places.

Enjoy Santorini, I'll be thinking of you two while I'm freezing my butt off in Joburg, sitting at the office!! I found April :)

Carry on enjoying, and I'll chat to you soon. Send April love from me too!! bye for now

:D hey you!! Ended up going to Bosnia for two days where there was no internet to be found can you believe it - although I was in this unbelievably tiny village.

Quite right, I DO have to give it back one day :) hopefully in the not too distant future!

It was great spending time with you too, I enjoyed every second :)

take care and chat soon. I'm going to upload some pics from the weekend just now!!

:) he he thanks luv!!!!

Name is married and lives in San Francisco, I don't really keep in touch with her, but my sister does. *Name*, a few years ago got divorced and went to live in Italy. Haven't seen or heard of her since :)

Please put some pics of your kids up, I'd so love to see them.

How long have you been in Sydney? I've heard it’s lovely. I've been to Perth which I adored, but that’s as far as I managed to go :)

Hey darlin

Whose christening was it? Please do send the Finedon crowd my regards and to your sister and *Name* of course.

The legalese is never easy. It does reduce all those emotions to a very cold and calculating level. I'm sorry you have to go through all of this – I wish I was there to hold your hand or offer you a shoulder.

I've seen his little flat – it's very small, but nice and clean and in a secure complex. I went shopping with him to help him chose kitchen stuff etc as men don't have a clue. I think you're right about the stark dose of reality biting. I don't think he realises how much living actually costs, and I think he overestimates his earning capacity. The phrase 'cut your pattern according to your cloth' springs to mind. I know he's also missing you a lot and still isn't sure he's done the right thing. Don't tell him I said this, but he was talking about your wedding day the other night, and actually started to cry. Shame. He needs to suffer (don't mean to be a bitch but he DOES).

Yeah the whole group pretty much feels the same way about *Name*. Remember when I told you how I was pissed off at her because of what she was doing to *Name*. You said you didn't understand why. For me it's all about integrity, and if she could do that to Name, she could do that to any one of us. How prophetic those words turned out to be. I just feel that she has no respect for anyone. The leopard skin is one example, *Name* is another. Even worse than doing these things, is the conspiracy of silence she weaves around them.

_____
_____

I won't be speaking to her again. Ever. *Name* feels the same way, as do *Name*, *Name*, blah blah blah. It's enough.

I'm so looking forward to seeing you – any arrival date is good for me – just let me know, and I'll be there to pick you up. Don't get a hire car, as you'll be staying with me and there's only parking for one. You may use my car, and I'll catch a lift to work with *Name*. On the evenings you'll be with your other friends, just use my car as well. If I need to go somewhere, we'll make a plan.

You WILL look stunning, you always do.

Enjoy your week my babe. I look forward to January.

Chat soon & lots of love

Hey my angel

I hope your weekend was fun and that you weren't too sad ☺

Just to fill you in on what's been happening in Joburg – I found out about *Name* sleeping with *Name* when I was in Prague. I'm really upset and disappointed with the betrayal, although I can't say I'm surprised. I don't think she respects anyone's boundaries where this kind of thing is concerned. I'm actually quite happy to write off the friendship – I don't need people who will do that, and secondly look you in the eye for months afterwards and lie about it. It shows a total lack of decency and a sad lack of conscience. Between the leopard skin, this, and a few other things, I've realised exactly what sort of a person and friend she is. Pretty much everyone in our group feels the same way, and does not want to associate with her anymore.

The weekend passed uneventfully – went out for dins with *Name* and *Name* last night – found this new place next to Bite called Sofia's which does tapas. Wow it was good. I can't wait to take you there for dinner when you get here. Speaking of that, have you decided on any dates etc?? I'm so looking forward to seeing you.

Please send everyone back home my regards. Lots of love to you my angel. Enjoy the week.

Yeeehaaaaa I'm so excited.

Yeah this is pretty much the same. It works out to 35000 including two nights in Beijing. Ideally, it would be great to spend four days there, and two extra in St Petersburg (we could get there on 23 Dec and have a white xmas)

Yes, let's get a comparative quote, never hurts to compare.

The only thing I'm concerned about, is that my dad's estate could take up to a year to wrap up, and they can't be more specific than that. I don't want to borrow from you not knowing exactly when I'll be able to pay you back. Perhaps in a few months the lawyer will have a better idea. I will go have a look and price tickets back from Beijing ☺

My dear *Name*

He is NOT the one for you. I think this was exactly what you needed to finally move on. Now you can put him in the past, and focus on meeting someone amazing.

I think now that you've had this closure, you can be happy and can move on.

You deserve much better than this jerk anyway, that I can promise you.

Chat soon, don't let this get you down.

Lots of love ☺ ☺ ☺

☺

Hello my dear friend

I'm glad you're happy. You still need to talk to him though it's important that you find out exactly what happened, and you deserve to know.

Either way, you will have closure. All I want if for you to be happy with some wonderful man.

Being back at work is horrible – I long for those warm, beautiful days in Dubrovnik.

Yes, that's my house. You will see it when you come to stay with me in Johannesburg ☺ ☺

Family are all good – all dying for these winter days to be over.

How are your mom and dad? Please send them my regards ☺

☺ I'm glad you saw him – let's hope he takes the initiative and contacts you, otherwise you call him ok???

Let me know how it goes. Lots of love, and chat soon

Hello *Name*

Apologies for not replying sooner, I have been away on business again.

I have a suggestion: How about planning to come over in our spring? September / October. It really is the loveliest of seasons here, and that way you could enjoy it twice in one year! Anyhow, whenever you decide to come, just let me know. I will do all I can to help you arrange things on this side too – as I'm sure you'll want to be seeing lots of people. Obviously, in Joburg I can take you where you need to go, but can also assist you with booking tickets and suchlike to other destinations in SA. It would be great to see you ☺

My birthday passed uneventfully – a nice enough day though ☺ with lots of love, calls, emails and attention. What could be better.

How are you doing? Are you fully recovered?

Sorry it’s been a while since my last message, wanted to drop you a line to say hi, and keep you abreast of our news in SA.

Both *Name* and I have recently moved - me back into my house in Parktown North, and *Name* and *Name* to a lovely house in Westcliff / Parktown. It’s really lovely, with an established garden, some amazing trees and a lot more space. The boys are also doing well, really cute. I babysat them for *Name’s* birthday on Tuesday.

I'm well too, have been invited to Dubrovnik, Croatia for a business trip in June, and am really looking forward to it. I will be extending my trip by a few days, to enjoy a bit of relaxing on the beach, and sightseeing.

Mom is also well, she's been in St Francis Bay for two weeks with Aunt *Name*.

Writer G
Hi *Name*

Sorry only getting back to you now. Had a really tough week as *Names*'s dad passed away last friday. We were at his bedside and he went peacefully but it was still very sad. We're all doing OK and were glad we could say goodbye.

Would love to get together next weekend. Let me know what day is better for you.

Take care,

Hi *Name*

Happy birthday for yesterday! It was also *Name*'s birthday! Hope you got spoilt and did something special to celebrate.

I'm writing to ask if I could put your name down as one of my references. I'm applying for a job in the Emirates and only want to tell everyone at The Lab if I succeed. I've been at The Lab for 10 years so I do feel it's time to move on and have a change but it won't be the end of the world if that doesn't happen just yet so I'd prefer to keep my options open there.

If you're OK with being my reference I'll need your address, email, tel(work), tel(home) and fax number.

How's *Name*? Are you guys EVER going to get married or do you prefer to be like Oprah and Stedman?! Just kidding!

Name and *Name* are doing well. *Name*'s already 15 months and talking more and more every day. She's also entering the tantrum stage which I'm dreading big time. No matter what you do people still look at you like it must be you...the mother!

Please say hi to *Name* and hope you enjoy your 31st year!
Take care,

Hi *Name*

Merry Christmas! Looks like you've been having a festive time. Are you guys around this week? Would love to get together if possible.

No problem, I really understand what it's like. We'll make a plan for soon.

Enjoy all the relatives!
Happy New Year!

Is that your birthday? So you're also an April baby - my little one is turning one on Friday so April babies are special. Can't believe a year's gone by.

Sorry I haven't written in a while...been quite busy and tired (dealing with teething baby). Hope you're doing well. Doing anything special for Easter?

Take care,

Hi *Name*

Great to hear you're coming next week. Would love to get together. My number is 072 083 8506.

Have a safe flight.
Love

Hi Sorry couldn't respond just now - at home with baby. The 9th is a monday isn't it? Aren't you only free in like 3weeks time?

Would love to join but evenings are just impossible at the moment. You'll have to take notes for me.

Thanks for asking,

In ways you wouldn't believe....it takes guilt to a whole new level but you also stop sweating the small stuff cos you just don't have the time or energy for it anymore. When you get back give me a call at The Lab as we need part-timers at the moment.

Hi *Name*

How's your week going? And how was the calm dude with Parkinsons?
Hope we can get together on Sat?

Have a Super Trooper day!

Hi

Sat will be fine for breakfast. Also wanted to forward the mail below to you re: Chinese New Year at Nanhua on Sunday. *Name* and *Name* are going and we were thinking of going too - want to join us?

Mama Mia....here I go again....bye bye.

Happy New Year to you too! I can see from all your facebook albums how busy you've been with weddings, dinners etc.

This Sat isn't good for us as *Name*'s working all day. What about Sunday? Lunch? We don't mind where so long as it's child friendly.

Hi *Name*

OK, next Sunday? Lunch at Fournos / Ciao Baby Bedford Centre?

How are your wedding plans coming along?

Hi there, have been thinking about you and wondering why we haven't got together in months. How did the op go? Must be tough for you having to do everything....I hope you haven't given him a bell to ring.

Have you set a wedding date yet?

Ja can't believe *Name* is almost 8 months. I want to wrap her in bubble wrap and keep her this age forever. Although think a person can only change so many nappies in their lifetime and my quota's fast running out.

Are you guys going away in Dec?

Hi *Name*

Thanks for your message. The wee one has bumped her head a few times now but decided we are probably born with more brain cells than we'll ever use so losing a few ain't gonna make a difference.

I enquired on your behalf about attending the Xmas lunch but unfortunately the powers that be were not in favour of this. Sorry, it would have been nice to have you there. Hope they do something good at Wits?!

Take care,

Hey *Name*,

How's it going? When's the due date? How are you both feeling? It's the most exciting and amazing thing that you'll ever experience....seeing your baby being born. *Name* is such a sweet name. Have you done up the baby room etc? I remember it took me ages to sort through all the clothes and categorise....I think women go a bit crazy when they

get into the nesting phase. I'm still in it and will probably be for a long time....so poor *Name* will just have to put up with me.

Name's doing great. She's started giggling and seems surprised she can even make this sound. It's so cute. It was really tough in the beginning 'cause the whole enormity of the situation just hit us the day we came home from the hospital....we were wondering how they could let us leave with her and so trustingly place her in our novice care. For the first few weeks she still looked a bit like an alien (not only in mama's tummy), maybe 'cause she was born a little prem at 37 weeks and only weighed 2.84KGs.

Anyhow....if you feel completely overwhelmed and strange for a while it's apparently normal.

It gets a whole lot better once they start smiling and cooing. Are you gonna take paternity leave? How many days do you get? How long is maternity leave in the UK? Here it's 4 months but I've taken 5 as I saved a lot of my leave from last year. Only going back to work in a month which I'm dreading big time but don't have much choice.

Who would ever think my whole mail to you would be about babies! Hope I haven't bored you to tears.

Hope all is well. Post the scan pics.

Yes would love to see more pics of *Name*. How are you coping? How's she sleeping? It's the most amazing scarey thing to experience hey? Coming to SA anytime soon? We can get together and discuss diaper changing techniques.

Name's getting cuter everyday. She's started rolling around in her cot and beaming at us in the morning. Can't really remember her at 2 months anymore unless I look at pics so cherish it 'cause it changes so quickly.

Big hug and kiss for your little cherub.

Hey, can't wait to see you in the Dec issue. What did you have to model?
Tell me more about your job - what do you have to do everyday and more important what is required to get further? (i.e. who do you have to sleep with / brown nose to get the ideal job?).

Books on Buddhism...'The Tibetan Book of Living and Dying' by Sogyal Rinpoche is a classic plus anything by the Dalai Lama. Like any religion / philosophy it's always good to read around to get a few perspectives before you can really get the essence of it.

Name's just recovering from an ear infection which was quite stressful. Otherwise she's getting cuter every day and spends a lot of time kicking, cooing, blowing raspberries

and giggling. I'm totally in love with her and still can't believe she came out of me! It's really the most amazing thing.

I agree about keeping things spicy and hot! Guys need to be kept on their toes.

Baby's calling, better go.

Love ya

Hi *Name*

Thanks for your mail and sorry I've taken a while to respond. *Name* was sick again the whole of last week with gastro this time. The anxiety and stress of having a sick child is overwhelming. I almost cracked. You just don't know what to do, how long it's gonna last and how to console the child.

So...coming to the first part of your mail, I think you're being realistic when it comes to having kids. I also struggled and continue to struggle with the centre-of- the- world-am-I syndrome. It's hard when you have something to complete and your baby DEMANDS your attention and guess who wins? So it's better waiting if you don't feel ready (not that one can ever really say they're 100% ready). It's also fine to decide not to have kids....the hard part is having to explain why to society. But for the record, it's also an amazing experience and I think you'd be a sweet and creative mom.

I'm also confused by *Name*'s behaviour towards you. This stems from the whole '*Name*' issue right? Is it because you continued to be his friend? Have you ever spoken to her about this? I know what you mean about the angry silences....silence is one of the most powerful manipulators of people I've come to realize. It feels like a form of torture and you're left to try fill in the blanks and make sense of it all.

You don't deserve to feel this Lucinda, no matter what happened to bring it about. At some point you may have to let go and stop hurting. I had to do that with *Name* and it's almost like she heard me 'cause that's when she contacted me but even if she hadn't I would have been OK. You've got to make peace with the situation and accept it in order to move on without anger, frustration or pain. You also need to remind yourself that whatever happened between you was not what you intended so that you release yourself from the whole cycle of pain.

I sound like an eccentric New Ager but it's probably because I'm also trying to get over losing *Name* (*Name*'s twin) and I also have to let go and accept it in order to move on.

Oh before I forget, could I have your tel number 'cause my cell was stolen at work about 2 weeks ago so lost all my numbers. I'm still on the same number.

Hope to chat soon,

Take care
Lots of love

P.S. Let me know when you appear in Marie Claire.

Hi *Name*

Sorry only getting back to you now. It's a pleasure and hope the dress fits. *Name* is between sizes at the moment so drowns in 6-12 months but looks a bit too worm like in 3-6 months.

She's doing much better thanks and is back to her kicking, cooing, giggling self. They just get cuter every day hey?

Hope to see you soon,
Love

Hi

Yes we love all beans and lentils so forward away! Back to work hasn't been easy but I'm slowly adjusting. *Name*'s at a good creche and is looked after by a paediatric nurse so I know she's in good hands but it still breaks my heart every time I leave her. How old is *Name* now? Are you working? Planning any more kids?

We're living in JHB and so are my parents and sister. How about your brother? What's he up to these days?

Take care

Hi,

Yes back to work which was really hard for the first week but now it's getting easier. I don't think you can ever say you have enough time for your kids. There's always something to feel guilty about as a parent I've come to realise. Are you working now? How are the kids?

Writer H
Oh ok it's working now ...but it wasn't earlier. Cool! :)
I hope everything goes well with your exams, just don't go out partying until the exams are over (that's something I would do to forget about the stress but it never did me any good).
Be good and good luck honey!

Hello (....then crawls under table)
I'm sorry, I know I owe you an email .....like I promised months ago. I suck :(

Hope you're well (not holding a grudge?)
How are things going on your side?

Hey I finally got my work permit, now I have to find a bloody job, oh crap ...what the hell was I complaining about when I didn't have the work permit, there was no pressure to do anything then.
Anyhoo I'm kinda excited too, I'd like to meet new people and have more cash to spend so this is good right.

Ok tell me what happened when you went to London, all that girl trouble thing. Have you heard from her since?
And how are things with you and your hot model GF?

I'm hardly online anymore, going through a faze I guess or that could also be called 'enjoying summer', it doesn't last very long here so I've got to make the most of it while I can hey.

When's the Depeche Mode gig on again? You going?

X

Hi *Name*

Sorry I haven't replied to your email. I've just been busy ...and not really on facebook that much these days.
Congrats on all your good fortuntes, you bloody deserve it buddy!
I hope this good streak carries on for you.

On my side, I finally got my work permit so that's good news for me too. I'm just a little nervous about going job hunting, especially the interviews. I'm out of practice with this whole working thing now.

Big hug to you and *Name*

So YOU'RE the Patient!! hahaha ...go figure, the psychologist needs his own head checked. :D
hugs you know I mean well don't you, I really hope it isn't serious.

I'm NOT skiving! I fuckin want to work ...I've never wanted to work so bad in my life! It's the immigration office taking their sweet time with my application, I think they've all gone into hybernation for the winter, they don't even answer their phones anymore!!

Married life is good, I highly reccommend it. Have you thought you might take the plunge then? :D
...and NO babies for me thank you, I really don't think I'm the Mummy type.

Calgary's still alright, weather's getting better so i've been out more too but now I'm thinking it might be good to move, maybe down to Vancouver. Or maybe I should just be patient.... blah

So it sounds like you're busy, yeah busy poking people on FB all day huh ;D
I'm looking forward to hearing some of the new songs with this new line-up, have you got new stuff out? I never go on MySpace anymore.

Always a good thing to live near a tube station. Highgate's nice too.

Good to hear from you, take care! XX

Hello ...hey we were busy! Actually didn't even hear the phone, TV was too loud ...*Name*'s fault completely :p
I heard your message this morning, I don't know if I should call you during working hours, I don't want you to get into trouble ("jeez this *Name* dude is always on the phone instead of working") ...hehehe

Do you really think we were going to miss the gig?! China?! No way!!
We were going to call you at ....ummm 9ish that evening for directions, we plan these things well ....hahahahha!

I'll call you later, thanks for getting us tickets :)
Woohoo we're already ready ....taking all our vitamins too

Ok I have to go, I'll talk to you later
Have a good day :D

Hehehehe ...cool so you remember me, *Names* sister!! :D
I thought ...well with the surname changed you might wonder who the hell this *Surname* chick is?! Just clarifying.

I'm immigrating, that's what I'm doing in Canada. Yeah it's my life long ambition to immigrate to a country that has temperatures colder than your average freezer. It's what I do for kicks you know, I punish myself in ways no one else would dream of ...all in the name of fun! Hehe

Ok seriously now, I got married in March this year and it seemed easier for me to move to Canada than for *Name* to move to the UK, he also had a cat at the time and he felt bad getting rid of it (anyway cat's dead now, no more excuses), maybe we'll move back to the UK after a few years.
All is good actually, I like it here and I'm happily married.

How about you, how are you doing?!
Are you still in Bolton? Don't you just miss those days at Anchor!? Ahhh

Anyhoo... glad to be in touch again, let me know how you've been keeping :)

Hi *Name*,
very good to hear from you! :D

Do you like it in Spain, can you speak Spanish fluently now??
Is *Name* the spanish girl I knew as *Name* (all this time I believed her name was *Name*), is she the same person from the Loop?
If it is her I'm so happy you are still together, she is such beautiful and kind person. Please tell her I say hello and send lots of hugs from me.
So when are the two of you going to get married? ;)

I'm having a great time in Canada, the people are very friendly and eveything is nice except for the weather. It's getting so cold now and I just want to stay in bed all day watching movies.
Oh I got married in March this year, my husband is Canadian. He's very good to me, I'm very lucky.

I'm still in touch with *Name*, she is living in Cape Town, South Africa, you can find her on my friends list if you want to contact her.

I wish I had kept in touch with more people but ....anyway I am glad *Name* is on Facebook, that is how I found you!!
I'm so happy *Name*'s going to be a daddy, he's going to be a fantastic father don't you think?

Take care! ***hugs***

Now the emails I sent to girls

Hi *Name* :D
No worries about the last email, lol I don't even remember it ...oh :(I'm sad you won't come see me in October but I understand. I'm sure it will be good visiting SA after 2.5 years. I bet you'll go back and everything will be the same ...haha that's what I find everytime I go back ...or maybe not?? Well enjoy it and let me know all about it.

Yes stay in touch, it's always good to hear from you and I'm sure EVENTUALLY we'll meet up.
Yip I need to get a job ASAP so I can go see you in SF too, I really want to go see what all the fussy is about, *Name* absolutely loved it there!

Hey best way to loose weight I find is to eat 20 portions (size of your fist) of fruit and veg a day for a week. Aim for 20 seriously, raw fruit and veg, cooked veg, veg soup, you name it. You won't feel hungry at all and when you do just have another piece of fruit or veg, just keep eating ...hahaha see if you can actually go through 20 portions, I never could!
Anyway good luck I hope that helps.
OMG I'm startnig to sound like the sham-wow guy! Yikes! lol

Lots of love and hugs!
Name

Hi *Name*
How's it going, it's great to hear from you ...and my apologies for taking my time to write back, I was going to email *Name* too ...but oops ...kinda left it too long.
How are you guys doing?

OMG if you hate the wind there then you're NOT going to like Calgary at all, wind = hate (the worst for me too) and it's windy here ALL THE BLOODY TIME!! Drives me mad!
It's dry here too, I get nose bleeds and my skin feels like it's going to crack.
Hahaha oh yes the weather is a big factor in life!

Name and I are looking at moving to Vancouver or somewhere in Vancouver Island, it's more like the UK I guess with the humidity (everything's green) ...but when is the real question here? I would go today if I could.

So does *Name* have to leave the UK? And if he goes does that mean you have to leave too?
Name said something about maybe checking out Canada, do it!! It's cool here (hahaha yes just after I had a little bitch about the weather) ...ok but don't come to Calgary cause

I might not be here, go to Vancouver.

There's the industrial, gothic scene here too. Im guessing it's bigger in Vancouver ...I'll find out, still haven't been there.
In Calgary the scene is tiny, this is more red-neck territory, they don't like anything too crazy here.

Anyway just get over here, let me know when and where, I'll help you guys out anyway I can.

Ciao for now
XX

Yes I'm so glad I found *Name*, I just typed in his name in the 'find a friend' search, it was that easy. He's an amazing guy.

I just saw some photos of you and *Name*, both of you look so happy ...I would say you ALMOST looks as happy as me and *Name* hehehehe
It great to see that :)

Oh all we need now is a cat ...just like Nero hehehehe. We really do need a cat though, I think there are mice in our ceilling, I'm dreading the day they come into our apartment.
Did I tell you I dated someone named Nero? lol

Did you cut your hair? Looking good!! ;)

I'll let you get back to your studying now ...take care. XX

Hello!

I'm so sorry you're on your own this week but hey look on the bright side, this gives you time to turn your house into a Santas Grotto while your Husband's gone. He'll be delighted ...lol

Ofcourse Fairlands in SA, I do know it.
Do you keep in touch with your Mom regularly, tell her I say hello, I wonder if she remembers us and our barbi dolls?
I'm sure your Mom misses you and her grandson, especially that he's so little, Granma's just love little babies. My Mom keeps pestering me to have kids but ...it's not going to happen.

We HAVE to keep in touch with my Mom. Since she got the knack of msn she's always online waiting for either my Sister or I to go online ...and heaven help us if we haven't signed into msn for a week! I admit it's great for her 3 hour video calls though or else the phone bills would be ridiculous.

I keep trying to sell the idea to my Sister that Canada is by far the best place in the world and that winter is actually the most exciting time of the year but she doesn't seem to be buying it. :D

I sent *Name* a link to your Uncles website, he's very interested, he had a quick look but he's too tired ...so am I for that matter.
I'm going to keep my ears and eyes open for any mention of dolfins now you know. Thanks

Ok that's me ....bedtime.
Take care and I hope your week goes quick so you get to see your hubby sooner.

Sands

That's ok *Name*, we understand but you did miss out on a great evening. How is *Name* doing now? Do you know what made her so sick?

About that sushi I'll have a word with *Name* and see what he says, I think it's a great idea, I'm having withdrawls too.

Appendix 4

Keyword analyses (1000-word level)

Writer A keywords (1000-word level)

N	Key word	Freq.	%	RC.	RC. %	Keyness	P
1	I	45	4.63	16	1.58	14.46	0.00014
2	NOT	17	1.75	1	0.10	13.24	0.00027
3	AM	8	0.82	0		6.44	0.01116
4	WAS	10	1.03	2	0.20	4.40	0.03596
5	WILL	8	0.82	1	0.10	4.27	0.03886
6	HAD	8	0.82	1	0.10	4.27	0.03886
7	MY	13	1.34	4	0.40	4.13	0.04209
8	AT	3	0.31	16	1.58	-7.17	0.00739

Writer B keywords (1000-word level)

N	Key word	Freq.	%	RC.	RC. %	Keyness	P
1	Y	8	0.79	0		6.18	0.01294
2	ON	12	1.19	2	0.20	5.86	0.01547
3	B	7	0.69	0		5.18	0.02278
4	YOU	40	3.97	23	2.27	4.26	0.03902
5	HEY	6	0.60	0		4.20	0.04044
6	LIKE	3	0.30	12	1.19	-4.27	0.03886

Writer C keywords (1000-word level)

N	Key word	Freq.	%	RC.	RC. %	Keyness	P
1	YA	8	0.81	0		6.35	0.01174
2	IM	7	0.71	0		5.33	0.02093
3	HEY	6	0.61	0		4.32	0.03762
4	NOT	8	0.81	1	0.10	4.19	0.04066
5	AND	17	1.73	36	3.56	-5.77	0.01627
6	AT	4	0.41	16	1.58	-5.80	0.01598

Writer D keywords (1000-word level)

N	Key word	Freq.	%	RC.	RC. %	Keyness	P
1	HEY	9	0.89	0		7.13	0.00755
2	YOU	45	4.44	23	2.27	6.69	0.00969
3	#	16	1.58	4	0.40	6.10	0.01352
4	COMPANY	6	0.59	0		4.17	0.04104
5	ON	10	0.99	2	0.20	4.10	0.04286
6	IS	5	0.49	15	1.48	-4.10	0.04289

Writer E keywords (1000-word level)

N	Key word	Freq.	%	RC.	RC. %	Keyness	P
1	UK	7	0.69	0		5.17	0.02302
2	HEE	6	0.59	0		4.18	0.04080
3	LIKE	3	0.30	12	1.19	-4.29	0.03832

Writer F keywords (1000-word level)

N	Key word	Freq.	%	RC.	RC. %	Keyness	P
1	I	44	4.36	16	1.58	12.60	0.00038
2	S	9	0.89	0		7.17	0.00742
3	T	7	0.69	0		5.18	0.02286
4	WAS	11	1.09	2	0.20	4.98	0.02565
5	DUBROVNIK	6	0.59	0		4.19	0.04056
6	ALSO	6	0.59	0		4.19	0.04056
7	AT	4	0.40	16	1.58	-6.08	0.01369

Writer G keywords (1000-word level)

N	Key word	Freq.	%	RC.	RC. %	Keyness	P
1	BABY	6	0.59	0		4.12	0.04249
2	WE	10	0.98	2	0.20	4.02	0.04500

Writer H keywords (1000-word level)

N	Key word	Freq.	%	RC.	RC. %	Keyness	P
1	I	38	3.78	16	1.58	8.54	0.00347
2	GOOD	15	1.49	3	0.30	6.86	0.00881
3	D	6	0.60	0		4.21	0.04009
4	ON	10	1.00	2	0.20	4.16	0.04146
5	AT	3	0.30	16	1.58	-7.57	0.00594
6	AND	14	1.39	36	3.56	-8.90	0.00285

Appendix 5

Keyword analyses (2000-word level)

Writer A/Writer X

N	Key word	Freq.	%	RC.	RC. %	Keyness	P
1	I	93	4.71	29	1.43	35.43	0.00000
2	AM	21	1.06	1	0.05	17.05	0.00003
3	WILL	25	1.27	3	0.15	16.47	0.00004
4	NOT	29	1.47	5	0.25	16.36	0.00005
5	WAS	16	0.81	3	0.15	7.96	0.00477
6	GOD	8	0.40	0		6.34	0.01180
7	MY	23	1.16	9	0.44	5.70	0.01692
8	LORD	7	0.35	0		5.33	0.02101
9	CHURCH	6	0.30	0		4.32	0.03772
10	OUT	10	0.51	2	0.10	4.30	0.03814
11	THAT	25	1.27	12	0.59	4.28	0.03854
12	ARE	10	0.51	25	1.23	-5.26	0.02187
13	LIKE	4	0.20	20	0.98	-9.01	0.00268
14	AT	4	0.20	28	1.38	-16.01	0.00006

Writer B/Writer X

N	Key word	Freq.	%	RC.	RC. %	Keyness	P
1	I	61	3.17	29	1.43	12.77	0.00035
2	HEY	9	0.47	0		7.59	0.00588
3	WAS	15	0.78	3	0.15	7.39	0.00656
4	YA	8	0.42	0		6.54	0.01055
5	MUST	7	0.36	0		5.50	0.01906
6	IM	7	0.36	0		5.50	0.01906
7	X	7	0.36	0		5.50	0.01906
8	OUT	10	0.52	2	0.10	4.50	0.03391
9	WINE	6	0.31	0		4.46	0.03470
10	OFF	8	0.42	1	0.05	4.35	0.03692
11	THE	43	2.24	70	3.44	-4.76	0.02907
12	AND	34	1.77	61	3.00	-5.89	0.01523
13	ARE	8	0.42	25	1.23	-6.96	0.00835
14	IT'S	3	0.16	16	0.79	-6.97	0.00831
15	AT	9	0.47	28	1.38	-7.86	0.00504
16	LIKE	3	0.16	20	0.98	-10.33	0.00131

Writer C/Writer X

N	Key word	Freq.	%	RC.	RC. %	Keyness	P
1	YOU	84	4.19	43	2.12	13.54	0.00023
2	B	13	0.65	0		11.27	0.00078
3	WAS	17	0.85	3	0.15	8.66	0.00325
4	HEY	10	0.50	0		8.24	0.00410
5	KINDA	8	0.40	0		6.23	0.01257
6	Y	8	0.40	0		6.23	0.01257
7	DAY	7	0.35	0		5.23	0.02220
8	ON	20	1.00	8	0.39	4.50	0.03398
9	I'LL	6	0.30	0		4.24	0.03953
10	A	55	2.74	36	1.77	3.88	0.04876
11	REALLY	4	0.20	15	0.74	-5.16	0.02311
12	LIKE	7	0.35	20	0.98	-5.22	0.02239
13	AT	12	0.60	28	1.38	-5.49	0.01914
14	ABOUT	3	0.15	19	0.93	-10.09	0.00149

Writer D/Writer X

N	Key word	Freq.	%	RC.	RC. %	Keyness	P
1	HEY	20	1.00	0		18.40	0.00001
2	YOU	82	4.09	43	2.12	12.45	0.00041
3	#	45	2.24	17	0.84	12.31	0.00045
4	ANYWAY	8	0.40	0		6.23	0.01254
5	IN	46	2.29	25	1.23	6.01	0.01422
6	THEN	10	0.50	1	0.05	5.94	0.01478
7	WELL	12	0.60	2	0.10	5.93	0.01490
8	BACK	12	0.60	2	0.10	5.93	0.01490
9	ON	22	1.10	8	0.39	5.85	0.01555
10	JO	7	0.35	0		5.23	0.02216
11	UP	16	0.80	5	0.25	4.92	0.02651
12	YEARS	12	0.60	3	0.15	4.39	0.03613
13	MARRIED	6	0.30	0		4.24	0.03947
14	YES	6	0.30	0		4.24	0.03947
15	SUNDAY	6	0.30	0		4.24	0.03947
16	LET	6	0.30	0		4.24	0.03947
17	COMPANY	6	0.30	0		4.24	0.03947
18	WOW	8	0.40	1	0.05	4.09	0.04313
19	WHEN	13	0.65	4	0.20	3.89	0.04860
20	YOUR	9	0.45	23	1.13	-5.15	0.02324
21	LIKE	4	0.20	20	0.98	-9.23	0.00237

Writer E/Writer X

N	Key word	Freq.	%	RC.	RC. %	Keyness	P
1	HEE	10	0.50	0		8.35	0.00385
2	I	53	2.67	29	1.43	7.20	0.00730
3	YOU	69	3.48	43	2.12	6.41	0.01132
4	SHE	10	0.50	1	0.05	6.04	0.01397
5	UK	7	0.35	0		5.31	0.02124
6	HUGS	7	0.35	0		5.31	0.02124
7	DAY	7	0.35	0		5.31	0.02124
8	WELL	11	0.55	2	0.10	5.15	0.02330
9	WAS	12	0.61	3	0.15	4.49	0.03409
10	IN	42	2.12	25	1.23	4.31	0.03787
11	DANIEL	6	0.30	0		4.30	0.03807
12	TO	40	2.02	64	3.15	-4.64	0.03123
13	ABOUT	6	0.30	19	0.93	-5.49	0.01907
14	LIKE	6	0.30	20	0.98	-6.22	0.01266
15	AT	11	0.55	28	1.38	-6.23	0.01259

Writer F/Writer X

N	Key word	Freq.	%	RC.	RC. %	Keyness	P
1	I	82	4.01	29	1.43	24.70	0.00000
2	S	21	1.03	0		19.02	0.00001
3	T	16	0.78	0		14.02	0.00018
4	YOU	77	3.76	43	2.12	9.13	0.00250
5	M	11	0.54	0		9.05	0.00262
6	DON	9	0.44	0		7.08	0.00781
7	WAS	15	0.73	3	0.15	6.68	0.00973
8	DUBROVNIK	8	0.39	0		6.09	0.01357
9	LL	7	0.34	0		5.11	0.02374
10	HEY	7	0.34	0		5.11	0.02374
11	ALSO	9	0.44	1	0.05	4.87	0.02736
12	LET	6	0.29	0		4.14	0.04185
13	SHE	8	0.39	1	0.05	3.97	0.04629
14	GOING	3	0.15	12	0.59	-4.33	0.03736
15	THE	45	2.20	70	3.44	-5.31	0.02117
16	#	4	0.20	17	0.84	-6.97	0.00828
17	AT	5	0.24	28	1.38	-14.93	0.00011

Writer G/Writer X

N	Key word	Freq.	%	RC.	RC. %	Keyness	P
1	ALSO	12	0.59	1	0.05	7.78	0.00527
2	MONTHS	7	0.35	0		5.19	0.02271
3	WHOLE	7	0.35	0		5.19	0.02271
4	DAY	7	0.35	0		5.19	0.02271
5	EVER	7	0.35	0		5.19	0.02271
6	TOGETHER	6	0.30	0		4.21	0.04030
7	SHE	8	0.40	1	0.05	4.05	0.04423

Writer H/Writer X

N	Key word	Freq.	%	RC.	RC. %	Keyness	P
1	I	81	4.02	29	1.43	24.77	0.00000
2	TOO	16	0.79	3	0.15	7.72	0.00545
3	GOOD	19	0.94	5	0.25	7.20	0.00728
4	GO	13	0.65	2	0.10	6.78	0.00921
5	D	8	0.40	0		6.20	0.01277
6	I'M	27	1.34	12	0.59	5.20	0.02257
7	I'LL	6	0.30	0		4.22	0.04000
8	TOUCH	6	0.30	0		4.22	0.04000
9	OH	6	0.30	0		4.22	0.04000
10	JUST	15	0.74	5	0.25	4.15	0.04160
11	ARE	11	0.55	25	1.23	-4.62	0.03158
12	AND	36	1.79	61	3.00	-5.87	0.01542
13	AT	6	0.30	28	1.38	-12.89	0.00032